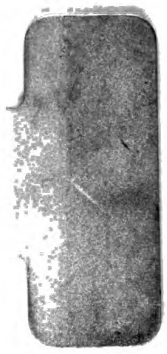
Lectures and Conferences on MATHEMATICAL STATISTICS AND PROBABILITY

NEYMAN



Lectures and Conferences
on
MATHEMATICAL STATISTICS AND PROBABILITY

By Jerzy Neyman

Professor and Director of Statistical Laboratory, University of California, Berkeley;
Formerly, Reader in Statistics, University College, London, and
Docent at the University of Warsaw, Poland

Second Edition, Revised and Enlarged

GRADUATE SCHOOL
U. S. DEPARTMENT OF AGRICULTURE
Washington: 1952

PRINTED IN THE UNITED STATES OF AMERICA

DEDICATION

This book is reverently and affectionately dedicated to the memory of my colleagues and friends lost during World War II. My association with them has contributed to the development of the ideas summarized in the following pages. In particular, I dedicate this book to the memory of:

ADAM HEIDEL, lost in a German concentration camp,
JANINA HOSIASSON, murdered by the Gestapo,
STANISŁAW KOŁODZIEJCZYK, missing,
KAZIMIERZ KORNIŁOWICZ, killed by a German bomb,
TADEUSZ MATUSZEWSKI, lost in a German concentration camp,
JAN PIEKAŁKIEWICZ, murdered by the Gestapo,
ANTONI PRZEBORSKI, starved during the German occupation of Warsaw,
JOZEF PRZYBOROWSKI, constrained to commit suicide when unable to escape the onrushing German armies,
STANISŁAW SAKS, murdered by the Gestapo,
HENRYK WILENSKI, missing.

J. Neyman

PREFACE

The original mimeographed edition (1938) of *Lectures and Conferences on Mathematical Statistics* was exhausted within two years of its publication. This, together with the subsequent continued inquiries from various persons and institutions, suggested that broad circles of statisticians are in need of a book such as this which gives the general ideas behind the theory of statistics and behind its applications. Unfortunately, certain circumstances prevented an earlier reissue of the book.

The present edition differs substantially from the first by an omission, by several additions and by reformulation of a considerable part of the earlier material. Owing to the extraordinary development of the econometric school on the one hand and of the works on stochastic processes on the other, the relevant Conference in the first edition became out of date and was omitted entirely. The interested reader is referred to articles in *Econometrica,* particularly to those of Ragnar Frisch, T. J. Koopmans, Oscar Lange and J. Marschak. In addition, he will find it both interesting and instructive to study the articles of J. L. Doob and W. Feller recently published in the Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability.[1]

Sporadic additions to the original material are inserted throughout the book. However, there are a few sections which deserve special mention. One such section is concerned with sampling human populations. Specifically, Parts 1 and 2 of Chapter III include a systematic presentation of the theory. Part 2 reproduces an article published some time ago in the *Journal of the American Statistical Association* and it is a pleasure to record my indebtedness to the Editor for the kind permission to do so.

The next substantial addition is Part 3 of Chapter III, which deals with spurious methods of studying correlation. Although the subject is not novel, the inclusion of a special section given to it seems justified by the fact that it appears to have been neglected by other authors while many empirical studies continue to involve errors of the kind described.

Although the earlier edition of *Lectures and Conferences* contains a counterpart of the present Chapter IV, there is a very substantial difference in presentation and a considerable addition of material. This chapter gives

[1] University of California Press, Berkeley and Los Angeles, 1949, 501 pp.

a three-cornered discussion of the ideas of estimation, from the point of view of Bayes' formula, from the point of view of confidence intervals and from the point of view of fiducial argument. Since the publication of the first edition of *Lectures and Conferences*, there has occurred a certain shift in "allegiances" exemplified by the fact that a large section devoted to fiducial distributions, present in an early edition of a book by an eminent author, does not appear in his subsequent books, which contain, instead, sections on confidence intervals. However, indications of the confusion of the Bayes' and the more modern treatment of the problem are still noticeable in certain sections of the literature and misconceptions involved in the fiducial argument appear about as frequently. For this reason it seemed advisable to subject the matter to a detailed discussion. Here I wish to record my hearty thanks to Professor E. S. Pearson, the Editor of *Biometrika*, for his kind permission to reproduce my article, originally published in that journal.

Part 4 of Chapter IV is entirely new and is given over to the brilliant recent result of Charles M. Stein.

Before concluding, I take pleasure in expressing my hearty thanks to Dr. Evelyn Fix for her invaluable help in the preparation of this book, for preparing the numerical illustrations, for reading and correcting the manuscript, and for kindly advice and suggestions.

J. NEYMAN
March, 1952

TABLE OF CONTENTS

CHAPTER I

The Modern Viewpoint on the Classical Theory of Probability and Its Applications. Tests of Statistical Hypotheses

(The contents of this chapter are based on three lectures delivered at the Graduate School of the United States Department of Agriculture in April, 1937.)

Introduction

After the original titles of my lectures had been fixed, I received a number of letters from members of the prospective audience and these letters forced me to modify the original programme and to place more emphasis than I had intended on concepts basic in the theory of probability and statistics.

The concept of probability has been discussed and defined in many different ways, each having its own advantage. It must be emphasized that, although the respective theories frequently contradict each other, this does not necessarily mean that some of them are wrong. Any theory is correct as long as the axioms on which it is based are not mutually contradictory and as long as there are no errors in deductions. Among the existing systems of axioms and theories deducible from them, we must make a choice. In this we shall be guided by considerations of usefulness or, by what frequently amounts to the same thing, our personal taste. It is important, however, to make clear the theory in which one is working. Otherwise, unnecessary misunderstandings may arise.

In my first lecture I shall describe the basic ideas of the theory of probability that I prefer and have had in mind when working on the theories of testing statistical hypotheses and of estimation.

So far as I am aware these views of mine are shared by E. S. Pearson and other workers attached to the Department of Statistics at University College, London. It may be, therefore, that the present lectures will help one to understand the whole of the work carried on in that centre.

It would be useless, of course, to try to develop the entire theory of probability in only two or three lectures. Therefore I shall concentrate on the general ideas, definitions, etc. Details of the theory of probability treated from the same point of view, though perhaps using different wordings, may be found in various books and papers, of which I shall mention the following:

1. H. Cramér: Random variables and probability distributions. Cambridge, 1937.
2. M. Fréchet: Recherches théoriques modernes sur la théorie des probabilités. Gauthier-Villars, Paris, 1937.
3. A. Kolmogoroff: Grundbegriffe der Wahrscheinlichkeitsrechnung. Julius Springer, Berlin, 1933.

Finally, an elementary systematic presentation is given in the recent book:

J. Neyman: First course in probability and statistics. Henry Holt and Co., New York, 1950.

The second lecture will be given entirely to the question of the possibility of applying the mathematical theory of probability to practical problems. The ideas developed here have grown out of reading such writers as E. Borel, L. v. Bortkiewicz, Karl Pearson and undoubtedly others but it is difficult to give exact references.

In the third and last lecture I shall deal with the somewhat narrower but still rather broad question of what is the meaning of a test of a statistical hypothesis and what are the grounds for choosing between several alternative tests. Material for the third lecture has been taken essentially from an article of mine which was published in 1929 in the *Proceedings* of the First Congress of Slavonic Mathematicians in Warsaw. The title of the article is "Méthodes nouvelles de vérification des hypothèses statistiques."

Part 1. On the Theory of Probability

1. Definition of probability. Probability as I shall define it will always refer to an object of a specified kind, say A, having a certain property, say B. Thus we may speak of the probability of a ball having the property of being black, of a person 36 years of age "having the property" of dying during the next twelve months, etc. It has been usual to define probability referring either to events or to propositions. Obviously the choice is very much a matter of convenience and it seems to me that speaking of the probabilities of objects having certain properties is convenient. Besides, it will be noticed that in this nomenclature we may speak also of probabilities of events. We will mean the probabilities of events having the property of actually occurring. Also it will be possible to speak of probabilities of propositions, which will mean the probabilities of propositions having the property of being true. The assumed system of expressions seems, therefore, to be not less general than the others.

In a mathematical definition, the actual wording used does not matter very much. However, it does have some importance since different wordings may appeal to intuition with different strengths and may give different emphases to the essential source of the concepts introduced. The essential

point in the concept of probability which I will use is that it will always refer to a specified *set* of objects, which I shall describe as the fundamental probability set. This point is emphasized in the wording adopted, since we agree to speak of the probability of a specified object A having a property B. It will be noticed that the process of specifying the object A is equivalent to specifying or perhaps even enumerating *all* objects that are "A" in distinction from those that are not. Now, all objects A will form what I shall call the *fundamental probability set* (F.P.S. for short). This will also be denoted by (A).[1]

It is obvious that in order to be able to enumerate all objects A, these objects must be well defined by a specification of one or more properties distinguishing the objects A from all other objects. This property will also be denoted by the same letter A.

Before proceeding further I shall explain the terms *logical sum* and *logical product* of two or more properties. Let B_1 and B_2 be any two properties. The property B_3 is a logical sum (or sum for short) of B_1 and B_2 if it consists in an object possessing *at least one* of the properties B_1 and B_2, and for this sum we shall write $B_3 = B_1 + B_2$. It will be convenient to use an expression like "an object $B_1 + B_2$" to denote an object possessing the property $B_1 + B_2$, etc.

A property B_4 will be called a logical product (or product for short) of the properties B_1 and B_2 if it consists in an object possessing *both* B_1 and B_2. We shall use the notation $B_4 = B_1B_2$ for this property and use the expression "an object B_1B_2" to denote an object possessing the property B_1B_2.

The above definitions are immediately extended to the sum and product of any number of properties, finite or infinite.

Turning now to the definition of probability of an object A possessing the property B, I want to emphasize that it requires the enumeration of *all* the objects A actually possessing the property B, i.e. all the objects possessing the property AB. According to the conventions already established, the set of those will be denoted by (AB).

Up to the present time our considerations have been perfectly general. Owing to the fact that the mathematical theory of sets is not commonly known, further steps leading to the definition of probability will have to be discussed twice, once on the assumption that the fundamental probability set (A) is finite and next, that it is anything, finite or infinite.

Suppose that the fundamental probability set (A) is finite, and denote by n the number of objects it contains. Further, let k be the number of

[1] "(x)" stands for "all x" and analogously for any letter in parentheses. This notation is in common use.

objects belonging to (A) and having the property B. The probability of an object A having the property B will be defined as the ratio k/n, and will be denoted by

$$P\{B \mid A\} = \frac{k}{n}. \tag{1}$$

In other words, the probability of an object A having the property B is defined as the proportion of objects A having the property B. The expression "the probability of an object A having a property B" is, of course, somewhat lengthy; we shall therefore use abbreviations such as "the probability of B," but it is necessary to remember the full meaning of these words.

Whenever there will be no danger of misunderstanding, the above notation can be simplified. For instance, if the probabilities that are calculated in the course of solving a certain problem refer always to the same fundamental probability set (A), the letter A may be omitted in the symbol of probability, whereupon $P\{B\}$ will suffice for $P\{B \mid A\}$. Sometimes, however, we shall have to deal not only with a fundamental probability set (A), but also with one or more others, each forming a part of (A). For instance, besides dealing with the probability of an object A having a certain property B', we might deal also with the probability of an object AB having the same property B' (or some other). In such cases the probabilities referring to objects A may be written without specifying their set, while probabilities referring to objects AB may not be: thus, $P\{B' \mid AB\}$ may be shortened to $P\{B' \mid B\}$, and $P\{B' \mid A\}$ may be shortened to $P\{B'\}$.

It is most important to distinguish the probabilities $P\{B' \mid A\}$ and $P\{B' \mid AB\}$. The former is the proportion of all objects A having the property B', while the latter is the proportion of objects AB having the property B' in addition to the property AB. Special care in distinguishing these two concepts is needed when we use shorter expressions and notations.

In order to emphasize this distinction we shall sometimes describe $P\{B' \mid A\}$ as the absolute probability of B' and $P\{B' \mid AB\}$ as the *relative* probability of B' given B. The relative probability of B' given B may or may not be equal to the absolute probability of B'. If it is, then we say that the property B' is *independent* of B.

It will be noticed that the definition of probability applies only to cases where the fundamental probability set is not empty, that is to say, only when it contains at least one element. Otherwise the word probability would have no meaning. It follows that whenever we speak of a probability, we imply that the fundamental probability set is not empty.

It follows from the definition that the probability P of any property, E, is a fraction between zero and unity. If $P = 0$, none of the elements of the F.P.S. has the property E. In this case we can conveniently describe

E as an *impossible* property. If on the other hand $P = 1$, it follows that the property E may be described as a *sure* property.[2] It is easily seen that the converses are true, namely that if E_1 and E_2 are an impossible and a sure property respectively, then $P\{E_1\} = 0$ and $P\{E_2\} = 1$. It will be noticed that the relative probability $P\{B' \mid B\}$ of B' given B has a definite meaning only if B is not an impossible property.

The characteristic feature of the above definition of probability is (i) that it refers to sets of objects and (ii) that it does not involve any reference to "equally probable" cases. In order to emphasize the consequences of the definition, I shall discuss a few examples.

Example 1.—A die has six faces, one and only one of which has six points on it. The probability of a side of the die having six points on it will be, according to our definition, always 1/6. No experiments with die throwing are able to alter this conclusion.

Example 2.—The probability of a side of the die having six points on it must be distinguished from the probability of getting six points on the die when the die is thrown.

Reading this last sentence once again and comparing it with the definition of probability, equation (1), one will easily see that, without further description of the situation, the definition of probability could not be applied to the throws. In speaking of "the probability of getting six points on the upper side of a die when throwing" and in trying to apply the definition of probability, we may have various things in mind.

(a) We may think of a set of 100 throws already carried out. Then there will be no difficulty in calculating the probability required.

(b) We may think of a set of some 100 future throws. In that case the probability required, say $P\{\text{six}\}$, will be just unknown. To establish its value, we should carry out the throws and count the cases with "six."

(c) Finally we may have in mind some hypothetical series of throws and discuss various probabilities referring to it. Usually such discussions consist in deducing values of one or more probabilities from the assumed hypothetical values of others. Some examples of such discussions will be found later.

Of the three ways of interpreting the ambiguously stated problem concerning the probability of getting "six" on a die when throwing, the last is the most fruitful. We shall see this a little further on when I shall speak of the so-called empirical law of large numbers.

Example 3.—Consider the familiar expansion $\pi = 3.14159 \cdots$ and denote by x_{1000} its thousandth decimal. What is the probability $P\{x_{1000} = 5\}$ of its being equal to 5? Here the question is not ambiguous and the answer

[2] "Sure property" is an English adaptation of the French phrase, "propriété certaine," as introduced by Maurice Fréchet and used in similar contexts.

is immediately found: the value of the probability $P\{x_{1000} = 5\}$ is actually unknown, but it is certainly either zero or unity. In fact, there is but one object satisfying the definition of x_{1000}. Therefore, the fundamental probability set consists of only one element and thus the denominator in the right hand side of equation (1) is equal to unity. The numerator may be equal to unity—this if x_{1000} is actually equal to 5—or to zero, if x_{1000} is not equal to 5. As the decimals in the expansion of π are known only to 707 places, x_{1000} is unknown and therefore we do not know whether $P\{x_{1000} = 5\}$ is zero or unity.

As I have mentioned before, probabilities may refer to some hypothetical probability sets, with assumed properties. This case is the one with which the theory is most often concerned, and is of extreme importance. Therefore I shall give two illustrations.

Example 4.—Consider a set F_1 of n die tosses, and denote by F_2 the set of $\frac{1}{2}n(n-1)$ different pairs that may be formed out of them, no element to be repeated in a pair. If certain properties of the set F_1 are given we may calculate the probability, say $P\{\text{six, six} \mid F_2\}$, of a pair of throws with two "sixes," referring it to F_2 as the F.P.S. The property of F_1 that is needed for the calculation of $P\{\text{six, six} \mid F_2\}$ consists in the probability $P\{\text{six} \mid F_1\}$ of getting a six in one throw. Assume, for instance, that

$$P\{\text{six} \mid F_1\} = \tfrac{1}{6}. \tag{2}$$

This would mean that among the n throws in F_1 there are exactly $n/6$ with six on the top face of the die, from which we could conclude that among the $\frac{1}{2}n(n-1)$ pairs of throws forming F_2 there are exactly

$$\frac{1}{12}n\left(\frac{1}{6}n - 1\right) = \frac{n(n-6)}{72} \tag{3}$$

such pairs that consist of two "sixes," and therefore that the probability

$$P\{\text{six, six} \mid F_2\} = \frac{n-6}{36(n-1)}. \tag{4}$$

It will be seen that the above result is purely hypothetical: *if* the connection between F_1 and F_2 is as described above, and *if* the probability of a specified property ("six") calculated with regard to F_1 is 1/6, *then* the probability $P\{\text{six, six} \mid F_2\} = (n-6)/36(n-1)$. Thus, *if* the probability set F_2 has the properties as specified in the conditions of the problem, *then* formula (4) holds good. We may notice at this stage that the properties of a probability set F_2 relevant for the calculation of probabilities may be given indirectly by specifying certain properties of some other set F_1 (or

of many other such sets), *and* by describing the connection between F_2 and F_1. A similar situation prevails in the following example.

Example 5.—Consider a series of n hypothetical experiments and assume that each of these experiments results either in an event E or in a failure to produce E, described as non-E. Assume further that a separate probability set is connected with each of the experiments, each set consisting of the same number m of elements and denote by F_i' the set corresponding to the ith experiment, $i = 1, 2, \cdots, n$. Suppose that whatever be i, the probability of the event E calculated with regard to F_i' is the same, that is,

$$P\{E \mid F_i'\} = p. \tag{5}$$

We may now consider still another probability set, say F_0, the elements of F_0 being all possible combinations of elements of the sets $F_1', F_2', \cdots, F_n'$ taken n at a time, where each element in the combination is selected from a different set. If each of the sets $F_1', F_2', \cdots, F_n'$ consists of the same number m of elements, then the set F_0 will consist of m^n elements.

The assumed properties of the sets $F_1', F_2', \cdots, F_n'$ and their connection with F_0 permit the calculation of various probabilities referring to F_0. For instance we may calculate the probability, say $P_{n,k}$, which frequently is picturesquely described as the probability of getting an event E exactly k times in the course of n independent trials, the probability of E in each trial being permanently equal to p. This probability is easy to calculate and is known to be equal to

$$P_{n,k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \tag{6}$$

But it is important to know what this formula denotes. This probability $P_{n,k}$ is no more and no less than the proportion of elements of the set F_0 that have the desired property of k "events" E and $n - k$ "events" non-E.

Again in this example, the calculation of the probability $P_{n,k}$ referring to the probability set F_0 was based on probabilities referring to the sets $F_1', F_2', \cdots, F_n'$ and on the structure of elements of F_0, each of them being composed of elements of $F_1', F_2', \cdots, F_n'$.

This is a typical situation and it will be convenient to introduce special terminology for its description. If the elements of any probability set F_0 are combinations of those of some other sets F_1, F_2, etc., then we shall say that the set F_0 is of a higher order than the sets F_1, F_2, $\cdots$. Thus we may distinguish probability sets of first, second, third, etc. order.

In Example 4 the set F_1 is of first, and the set F_2 of second order. In Example 5 the sets $F_1', F_2', \cdots, F_n'$ are of first order and the set F_0 of the second. It is easy to construct examples in which there will be probability sets of three or more successive orders.

In what I have just said I used the expressions "experiments," "results," "events," which were not directly involved in the definition of probability. I want to emphasize that these expressions are no more than a picturesque description of fundamental probability sets and that if purity of language really were demanded, they should not be used. However, these and similar expressions are very frequent in all works on probability. They were established in olden days when the point of view regarding probability theory was somewhat different. We hold on to them now because of their convenience. This point will be discussed later when I shall speak of applications and of the law of large numbers.

We shall notice now that a description of a conceptual experiment, as in the above examples, amounts really to a description of probability sets. As the sets were classified, so will be classified the corresponding hypothetical experiments. Therefore we shall speak of experiments of the first, second, third, $\cdots$ order.

In order to clear away any possible misunderstanding, let us consider again the probability sets involved in the last two examples, and illustrate them graphically. The set F_1 of Example 4 may be represented by the use of the letter s for "six," and the letter r for "not-six." With $n = 12$, we might have the following picture:

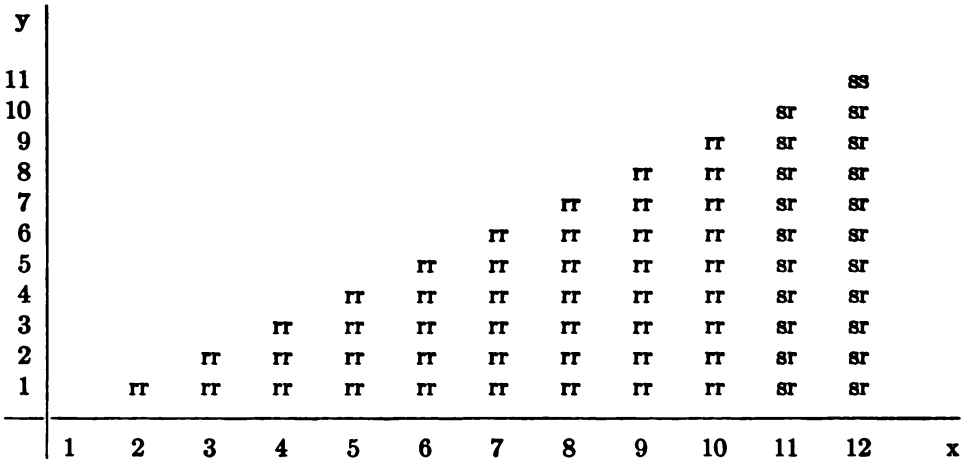
r	r	r	r	r	r	r	r	r	r	s	s
1	2	3	4	5	6	7	8	9	10	11	12

The numbers 1 to 12 below the line represent the ordinal numbers of the elements of F_1.

To represent F_2 diagrammatically it will be convenient to use two dimensions. Each element of F_2 is represented by rr, rs, sr, or ss. The rectangular coordinates x and y of an element of F_2 are equal to the ordinal numbers of the two elements of F_1 making up this element of F_2. As x can never be equal to y, i.e., no element of F_1 is to be repeated, it is permissible to take $x > y$. There will be only one element of F_2 possessing the property "six-six" (ss), that composed of the eleventh and twelfth elements of F_1. It may be seen from Figure 1 that the number of elements forming F_2 is 66 and that, therefore, $P\{\text{six, six} \mid F_2\} = 1/66$, which agrees perfectly with formula (4) above, if n therein be set equal to 12.

We may now illustrate the connection between the probability sets F_0 and F_1', F_2', $\cdots$, F_n' of Example 5. Let us put $k = n = 2$, $m = 6$, $p = 1/6$, so that among the six elements forming either F_1' or F_2' there will be only one possessing the property E, the other five, denoted by G, being non-E. Let E in both sets be the 6th element. Any element of F_0 is formed by combining an element of F_1' with some element of F_2'. Therefore, it will be convenient to represent each element of F_0 by a point on a plane whose

FIGURE 1



coordinates x and y are equal to the ordinal numbers of the elements of F_1' and F_2', the combination of which produces the element of F_0 under consideration (see Figure 2). All the elements of F_0 possess the required property of being composed of elements of F_1' and F_2', but only one of the 36 is EE. The resulting probability $P_{2,2} = 1/36$ is in agreement with the binomial formula (6).

FIGURE 2

y						
6	GE	GE	GE	GE	GE	EE
5	GG	GG	GG	GG	GG	EG
4	GG	GG	GG	GG	GG	EG
3	GG	GG	GG	GG	GG	EG
2	GG	GG	GG	GG	GG	EG
1	GG	GG	GG	GG	GG	EG
	1	2	3	4	5	6 x

I hope that it is not necessary to insist that the above results, namely,

$$P\{EE \mid F_2\} = P\{\text{six, six} \mid F_2\} = \tfrac{1}{36} \quad \text{(Ex. 4)} \tag{7}$$

and

$$P\{EE \mid F_0\} = P\{\text{six, six} \mid F_0\} = \tfrac{1}{36} \quad \text{(Ex. 5)} \tag{8}$$

do not represent any sort of paradox. Both probabilities are calculated correctly and they differ only because they refer to different probability sets, F_2 and F_0. This emphasizes the fact that probabilities refer to prob-

ability sets and that failure to specify the probability set properly may, and usually does, cause misunderstanding.

Example 6.—The inclusion of the present example is occasioned by certain statements of Harold Jeffreys [3] which suggest that, in spite of my insistence on the phrase, "probability that an object A will possess the property B," and in spite of the five foregoing examples, the definition of probability given above may be misunderstood.

Jeffreys is an important proponent of the subjective theory of probability designed to measure the "degree of reasonable belief." His ideas on the subject are quite radical. He claims [4] that no consistent theory of probability is possible without the basic notion of degrees of reasonable belief. His further contention is that proponents of theories of probabilities alternative to his own forget their definitions "before the ink is dry." [5] In Jeffreys' opinion, they use the notion of reasonable belief without ever noticing that they are using it and, by so doing, contradict the principles which they have laid down at the outset.

The necessity of any given axiom in a mathematical theory is something which is subject to proof. For example, it was possible to prove that many of the theorems taught for decades in calculus depend on the famous axiom of Zermelo which by itself seems very doubtful to many mathematicians. The method of proof is as follows: One assumes that a given theorem is true and then deduces that the axiom subject to doubt must be true also.

However, Dr. Jeffreys' contention that the notion of degrees of reasonable belief and his Axiom 1 [6] are necessary for the development of the theory of probability is not backed by any attempt at proof. Instead, he considers definitions of probability alternative to his own and attempts to show by example that, if these definitions are adhered to, the results of their application would be totally unreasonable and unacceptable to anyone. Some of the examples are striking. On page 300, Jeffreys refers to an article of mine [7] in which probability is defined exactly as it is in the present volume. Jeffreys writes:

> The first definition is sometimes called the "classical" one, and is stated in much modern work, notably that of J. Neyman.

[3] Harold Jeffreys: Theory of probability. Clarendon Press, Oxford, 1939, vi + 380 pp.

[4] Jeffreys, *op. cit.*, p. 300.

[5] Jeffreys, *op. cit.*, p. 303.

[6] "Given p, q is either more or less probable than r, or both are equally probable; and no two of these alternatives can be true." Jeffreys, *op. cit.*, p. 16.

[7] J. Neyman: "Outline of a theory of statistical estimation based on the classical theory of probability." *Phil. Trans. Roy. Soc. London,* Ser. A, Vol. 236 (1937), pp. 333–380.

However, Jeffreys does not quote the definition that I use but chooses to reword it as follows:

> If there are n possible alternatives, for m of which p is true, then the probability of p is defined to be m/n.

He goes on to say:

> The first definition appears at the beginning of De Moivre's book (*Doctrine of Chances,* 1738). It often gives a definite value to a probability; the trouble is that the value is one that its user immediately rejects. Thus suppose that we are considering two boxes, one containing one white and one black ball, and the other one white and two black. A box is to be selected at random and then a ball at random from that box. What is the probability that the ball will be white? There are five balls, two of which are white. Therefore, according to the definition, the probability is 2/5. But most statistical writers, including, I think, most of those that professedly accept the definition, would give $(\frac{1}{2})\cdot(\frac{1}{2}) + (\frac{1}{2})\cdot(\frac{1}{3}) = \frac{5}{12}$. This follows at once on the present theory, the terms representing two applications of the product rule to give the probability of drawing each of the two white balls. These are then added by the addition rule. But the proposition cannot be expressed as the disjunction of five alternatives out of twelve. My attention was called to this point by Miss J. Hosiasson.

The solution, 2/5, suggested by Jeffreys as the result of an allegedly strict application of my definition of probability is obviously wrong. The mistake seems to be due to Jeffreys' apparently harmless rewording of the definition. If we adhere to the original wording and, in particular, to the phrase "probability of an object A having the property B," then, prior to attempting a solution, we would probably ask ourselves the questions: "What are the 'objects A' in this particular case?" and "What is the 'property B,' the probability of which it is desired to compute?" Once these questions have been asked, the answer to them usually follows and determines the solution.

In the particular example of Dr. Jeffreys, the objects A are obviously not balls, but pairs of random selections, the first of a box and the second of a ball. If we like to state the problem without dangerous abbreviations, the probability sought is that of a pair of selections ending with a white ball. All the conditions of there being two boxes, the first with two balls only and the second with three, etc., must be interpreted as picturesque descriptions of the F.P.S. of pairs of selections. The elements of this set fall into four categories, conveniently described by pairs of symbols $(1,w)$, $(1,b)$, $(2,w)$, $(2,b)$, so that, for example, $(2,w)$ stands for a pair of selections in which the second box was selected in the first instance, and then this was followed by the selection of the white ball. Denote by $n_{1,w}$, $n_{1,b}$, $n_{2,w}$ and $n_{2,b}$ the (unknown) numbers of the elements of F.P.S.

belonging to each of the above categories, and by n their sum. Then the probability sought is

$$P\{w \mid \text{pair of selections}\} = \frac{n_{1,w} + n_{2,w}}{n}. \tag{9}$$

The conditions of the problem imply

$$P\{1 \mid \text{pair of selections}\} = \frac{n_{1,w} + n_{1,b}}{n} = \frac{1}{2}, \tag{10}$$

$$P\{2 \mid \text{pair of selections}\} = \frac{n_{2,w} + n_{2,b}}{n} = \frac{1}{2}, \tag{11}$$

$$P\{w \mid \text{pair of selections beginning with box No. 1}\} = \frac{n_{1,w}}{n_{1,w} + n_{1,b}} = \frac{1}{2}, \tag{12}$$

$$P\{w \mid \text{pair of selections beginning with box No. 2}\} = \frac{n_{2,w}}{n_{2,w} + n_{2,b}} = \frac{1}{3}. \tag{13}$$

It follows

$$n_{1,w} = \tfrac{1}{2}(n_{1,w} + n_{1,b}) = \tfrac{1}{4}n, \tag{14}$$

$$n_{2,w} = \tfrac{1}{3}(n_{2,w} + n_{2,b}) = \tfrac{1}{6}n, \tag{15}$$

$$P\{w \mid \text{pair of selections}\} = \tfrac{5}{12}. \tag{16}$$

The method of computing probability used here is a direct enumeration of elements of the F.P.S. For this reason it is called the "direct method." As we can see from this particular example, the direct method is occasionally cumbersome and the correct solution is more easily reached through the application of certain theorems basic in the theory of probability. These theorems, the addition theorem and the multiplication theorem, are very easy to apply, with the result that students frequently manage to learn the machinery of application without understanding the theorems. To check whether or not a student does understand the theorems, it is advisable to ask him to solve problems by the direct method. If he cannot, then he does not understand what he is doing.

Checks of this kind were part of the regular program of instruction in Warsaw where Miss Hosiasson was one of my assistants. Miss Hosiasson was a very talented lady who has written several interesting contributions to the theory of probability. One of these papers [8] deals specifically with

[8] Janina Hosiasson: "Quelques remarques sur la dépendance des probabilités a posteriori de celles a priori." *C.R., Premier Congres des Math. des Pays Slaves,* Warszawa, 1929, pp. 375–382.

various misunderstandings which, under the high sounding name of paradoxes, still litter the scientific books and journals. Most of these paradoxes originate from lack of precision in stating the conditions of the problems studied. In these circumstances, it is most unlikely that Miss Hosiasson could fail in the application of the direct method to a simple problem like the one described by Dr. Jeffreys. On the other hand, I can well imagine Miss Hosiasson making a somewhat mischievous joke.

Some of the paradoxes solved by Miss Hosiasson are quite amusing. The facility with which one is able to resolve these paradoxes may serve as a test as to whether or not the definition of probability is properly understood. The following paradox is taken from the "Treatise on Probability" by J. M. Keynes (London, 1921, p. 378). Like Dr. Jeffreys, Lord Keynes was also a proponent of the subjective theory of probability.

Consider an urn U of which it is known that it contains exactly n balls. About the color of the balls no information is available. Denote by m the number of black balls in the urn. Because of the complete lack of information as to the color of the balls and since there are $n+1$ possible hypotheses about the value of m, namely $m = 0, 1, 2, \cdots, n$, the subjective theory of probability ascribes to each of these hypotheses the same probability, namely $1/(n+1)$. Granting this, it is easy to show that the probability, say $P(B)$ that a ball drawn from the urn will be black is $P(B) = \frac{1}{2}$. This conclusion, by itself, is not questioned. However, Lord Keynes seems to have been puzzled by the circumstance that what applies to black balls should equally apply to white balls and yellow balls. Therefore, if we denote by $P\{W\}$ and $P\{Y\}$ the probabilities that the ball drawn will be white and that it will be yellow, respectively, then $P\{W\} = P\{Y\} = P\{B\} = \frac{1}{2}$.

Further, since the colors white, yellow and black are exclusive, the probability that the ball drawn will be either black, white or yellow would appear to have the absurd value $P\{B + W + Y\} = 1.5$. How come? The reader may wish to try to resolve this "paradox" on his own. If he does not succeed, then he may find it interesting to consult the paper of Miss Hosiasson.

2. More general definition of probability. The foregoing definitions and examples are perhaps sufficient to explain the basic ideas underlying the theory of probability when the fundamental probability set is finite. Let us now turn to the more general case and assume that the F.P.S., say (A), is anything, finite or infinite. As formerly, let us denote by (B) the set of elements of (A) that have some distinctive property B.

The definition of probability I am going to give will apply only to certain sets (A) and to certain properties B, not to all possible ones. In fact, we shall require that the following postulates be satisfied by the class of

subsets (B) of A which correspond to the properties B for which the probability will be defined. This class will be denoted by $((B))$.

It will be assumed

(1) that the class $((B))$ includes (A) so that (A) is an element of $((B))$.

(2) that for the class $((B))$ it is possible to define a single-valued function $m(B)$, called the *measure* of (B), wherefore the sets (B) belonging to the class $((B))$ will be called measurable. The assumed properties of the measure are as follows:

(a) Whatever be (B) of the class $((B))$, $m(B) \geqq 0$.

(b) If (B) is empty (does not contain any element at all), then it is measurable and $m(B) = 0$.

(c) The measure of (A) is greater than zero.

(d) If (B_1), (B_2), $\cdots$, (B_n), $\cdots$ is any at most denumerable set of measurable subsets, then their sum, (ΣB_i), is also measurable. If no two subsets (B_i) and (B_j) (where $i \neq j$), have common elements, then $m(\Sigma B_i) = \Sigma m(B_i)$.

(e) If (B) is measurable, then the set $(\bar{B})$ of objects A not possessing the property B is also measurable and consequently, owing to (d), $m(B) + m(\bar{B}) = m(A)$.

Under the above conditions the probability, $P\{B \mid A\}$ of an object A having the property B will be defined as the ratio

$$P\{B \mid A\} = \frac{m(B)}{m(A)}.$$

The probability $P\{B \mid A\}$, or $P\{B\}$ for short, may be called the absolute probability of the property B. Denote by B_1B_2 the property of A consisting in the presence of both B_1 and B_2. It is easy to show that if (B_1) and (B_2) are both measurable, then (B_1B_2) will be measurable also. If $m(B_2) > 0$ then the ratio, say

$$P\{B_1 \mid B_2\} = \frac{m(B_1B_2)}{m(B_2)},$$

will be called the relative probability of B_1 given B_2. This definition of the relative probability applies when the measure $m(B_2)$ as defined for the fundamental probability set (A) is not equal to zero. If, however, $m(B_2) = 0$, but we are able to define some other measure, say m', applicable to (B_2) and to a class of its subsets including (B_1B_2) such that $m'(B_2) > 0$, then the relative probability of B_1 given B_2 will be defined as

$$P\{B_1 \mid B_2\} = \frac{m'(B_1B_2)}{m'(B_2)}.$$

Whatever may be the case we shall have

$$P\{B_1B_2\} = P\{B_1\}P\{B_2 \mid B_1\} = P\{B_2\}P\{B_1 \mid B_2\}. \tag{17}$$

It is easy to see that if the fundamental probability set is finite, then the number of elements in any of its subsets will satisfy the definition of measure. On the other hand, if (A) is the set of points filling up a certain region in n-dimensional space, then the measures of Borel and of Lebesgue will satisfy the definition used here.

If the objects A are not points (e.g., if they are certain lines, etc.), the above definition of probability can still be applied, provided it is possible to define a measure over a class of subsets of (A). One way of achieving this, which is frequently applicable, is to establish a one-to-one correspondence between the objects of (A) and some other objects (A') for which a measure has already been defined. If (B') is any measurable subset of (A') and (B) the corresponding subset of (A), then the measure of (B) can be defined to be equal to that of (B').

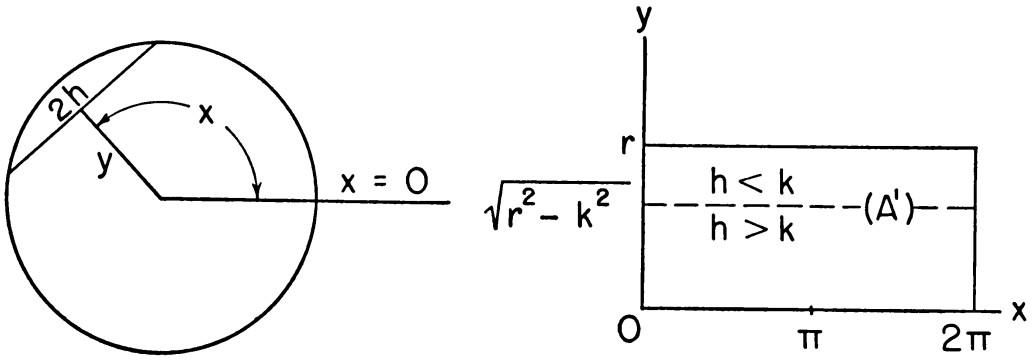
If a one-to-one correspondence between (A) and (A') can be established at all, then it usually will be easy to establish it in more than one way and each definition of correspondence between objects A and objects A' will imply, or as one occasionally says, induce a new definition of measure for subsets of (A). This, for instance, is the case when the objects A are chords in a circle C of radius r and objects A' points in a plane. It may be useful to consider two of the possible ways of establishing a one-to-one correspondence between the chords and the points leading to two different definitions of measure of the subsets of chords. Specifically, we will discuss the so-called Bertrand's problem which consists in determining the probability that a chord drawn "at random" in the circle C will have its length $2h$ greater than some specified value $2k < 2r$.

(i) Denote by x the angle between a fixed direction and the radius perpendicular to any given chord A, in a circle of radius r. Further, let y be the perpendicular distance of the chord A from the centre of the circle C. Now let A' denote a point on the xy plane with coordinates x and y; then there will be a one-to-one correspondence between the chords (A) of length $0 \leqq 2h \leqq 2r$ and the points of a rectangle, say (A'), defined by two pairs of conditions $[(0 \leqq x < \pi)(0 \leqq y \leqq r)]$ and $[(\pi \leqq x < 2\pi)(0 < y \leqq r)]$. The class of measurable subsets of chords may now be defined to be composed of all such subsets which correspond to subsets of (A') that are measurable in the sense of Borel. This includes the subset (AB) of chords

with lengths $2h > 2k$. In fact, these chords correspond to points, say $A'B'$ in (A') with their coordinate $y < \sqrt{r^2 - k^2}$. The set of points $(A'B')$ fills in a rectangle (apart from some points on the boundary) and its Borel measure is equal to the area of this rectangle, namely $2\pi\sqrt{r^2 - k^2}$. It follows that the probability in which we are interested is $P\{h > k\} = \sqrt{1 - (k/r)^2}$.

(ii) Denote by x and y the angles between a fixed direction and the radii pointing towards the two ends of a given chord A. If A'' denotes a point on a plane with coordinates x and y, then there exists a one-to-one corre-

FIGURE 3



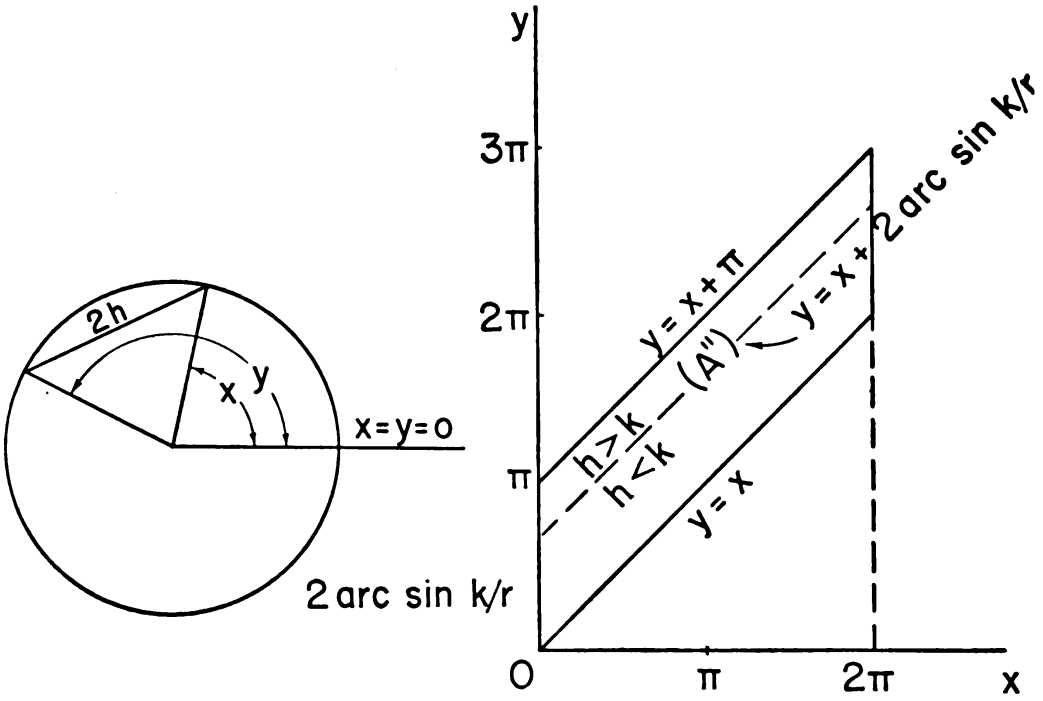
Solution 1.—Here the set (A) of chords is mapped on the rectangle (A'), the correspondence between chords and points in (A') being a one-to-one correspondence.

spondence between the chords of the set (A) and the points within the parallelogram (A'') (see Figure 4) determined by the two pairs of conditions $[(0 \leqq x < \pi)(x \leqq y \leqq x + \pi)]$ and $[(\pi \leqq x < 2\pi)(x \leqq y < x + \pi)]$. If (A_1'') is a subset of (A'') which is measurable in the sense of Borel and if (A_1) is the corresponding subset of chords, then define (A_1) as measurable and let the measure $m(A_1)$ be equal to the Borel measure of (A_1''). The points in (A'') which correspond to chords with lengths exceeding $2k$ lie above the dotted line $y = x + 2$ arc sin k/r. Since these points fill in a parallelogram, the set is measurable and its Borel measure coincides with the area of the parallelogram, namely $2\pi(\pi - 2 \text{ arc sin } k/r)$. Since the measure of the entire set (A) is equal to that of the entire set (A'') which is $2\pi^2$, it follows that the probability $P\{h > k\} = 1 - (2/\pi)$ arc sin k/r.

It is seen that the two solutions differ and it may be asked which of them is correct. The answer is that both are correct, but that they correspond to different conditions of the problem. In fact the question "what is the probability of a chord having its length greater than $2k$" does not specify the problem entirely. This problem is only determined when we define the measure appropriate to the set (A) and the subsets of (A) to be considered. We may describe this differently, using the terms "random

experiments" and "their results." We may say that to have a problem of probability determined, it is necessary to define the method by which the randomness of an experiment is attained. Describing the conditions of the problem concerning the length of a chord that lead to the first solution (Figure 3), we could say that in selecting at random a chord A, we first pick at random the direction of a radius, all directions being "equally

FIGURE 4



Solution 2.—Here the set (A) of chords is mapped on the parallelogram (A''), the correspondence between chords and points in (A'') being a one-to-one correspondence.

probable," and then, also at random, we select the distance between the centre of the circle and the chord, all values between zero and r being "equally probable." It is easy to see what would be the description in the same language of the random experiment leading to the second solution (Figure 4).

We frequently use this way of speaking, but it is necessary to remember that behind such words, as e.g., "picking at random a direction, all of them being equally probable," there is a definition of the measure appropriate to the fundamental probability set and its subsets. I want to emphasize that in all my writings a phrase like the previous one in quotation marks is no more than a way of describing the fundamental probability set and its appropriate measure. The concept "equally probable" is not in any way

involved in the definition of probability adopted and it is a pure convention that the statement

"In picking a chord at random, we first select a direction, all directions being equally probable; and then we choose a distance between the centre of the circle and the chord, all values of the distance between zero and r being equally probable."	Means no more and no less than	"For the purpose of calculating the probabilities concerning chords in a circle, the measure of any set (A) of chords is defined as that of the set (A') of points, each with coordinates x and y and such that for any chord A in (A), x is the direction of the radius perpendicular to A and y the distance of A from the centre of the circle. (A) is measurable only if (A') is so."

However free we are in mathematical work to use words that we find convenient as long as they are clearly defined, our choice must be justified in one way or another. The justification for speaking of the definition of measure within the fundamental probability set in terms of imaginary random experiments lies in the empirical fact which Bortkiewicz [9] insisted upon calling the "law of large numbers." This law says that, given a purely mathematical definition of a probability set including the appropriate measure, we are able to construct a real experiment, possible to carry out in any laboratory, with a certain range of possible results and such that if it is repeated many times, the relative frequencies of these results and their different combinations in small series approach closely the values of probabilities as calculated from the definition of the fundamental probability set. Examples of such real random experiments are provided by the experience of roulette,[10] by the experiment of throwing a needle [11] so as to obtain an analogy to the problem of Buffon, and by various sampling experiments based on Tippett's random numbers.[12]

These examples show that random experiments corresponding in the sense described to mathematically defined probability sets are possible. However, frequently they are technically difficult. E.g., if we take any coin and toss it many times, it is very probable that the frequency of heads will not approach 1/2. To get this result we must select what could be called a well-balanced coin and we must work out an appropriate method

[9] L. von Bortkiewicz: *Die Iterationen.* Julius Springer, Berlin, 1917, x + 205 pp.

[10] Bortkiewicz, *loc. cit.*

[11] This is mentioned by É. Borel, *Éléments de la Théorie des Probabilités,* Hermann, Paris, 1909, vii + 205 pp. Cf. p. 106.

[12] L. H. C. Tippett: "Random sampling numbers." *Tracts for Computers,* No. XV, Cambridge University Press, 1927, viii + 26 pp.

of tossing. Whenever we succeed in arranging the technique of a random experiment, such that the relative frequencies of its different results in long series approach, sufficiently in our opinion, the probabilities calculated from a fundamental probability set (A), we shall say that the set adequately represents the method of carrying out the experiment.

We shall now draw a few obvious but important conclusions from the definition of probability which we have adopted.

(1) If the fundamental probability set consists of only one element, any probability calculated with regard to this set must have the value either zero or unity.

(2) If all elements of the fundamental probability set (A) possess a certain property B_0, then the absolute probability of B_0, and also its relative probability, given any other property B_1, must be equal to unity, so that $P\{B_0 \mid A\} \equiv P\{B_0\} = P\{B_0 \mid B_1\} = 1$. On the other hand, if it is known only that $P\{B_0\} = 1$, then it does not necessarily follow that $P\{B_0 \mid B_1\}$ must be equal to unity.

3. Random variables. We may now proceed to the definition of a random variable. We shall say that x is a random variable if it is a single-valued measurable function (not a constant) defined within the fundamental probability set (A) with the exception perhaps of a set of elements of measure zero. We shall consider only cases where x is a real numerical function. If x is a random variable, then its value corresponding to any given element A of (A) may be considered as a property of A, and whatever the real numbers $a < b$, the definition of (A) will allow the calculation of the probability, say $P\{a \leqq x < b\}$ of x having a value such that $a \leqq x < b$.

We notice also that as x is not constant in (A), it is possible to find at least one pair of elements, A_1 and A_2 of (A), such that the corresponding values of x, say $x_1 < x_2$ are different. If we denote by B the property distinguishing both A_1 and A_2 from all other elements of (A), and if $a < b$ are two numbers such that $a < x_1 < b < x_2$, then $P\{a \leqq x < b \mid B\} = \frac{1}{2}$. It follows that if x is a random variable in the sense of the above definition, then there must exist such properties B and such numbers $a < b$ that $0 < P\{a \leqq x < b \mid B\} < 1$.

It is obvious that the above two properties are equivalent to the definition of a random variable. In fact, if x has the properties (a) that whatever $a < b$ the definition of the fundamental probability set (A) allows the calculation of the probability $P\{a \leqq x < b\}$, and (b) that there are such properties B and such numbers $a < b$ that $0 < P\{a \leqq x < b \mid B\} < 1$, then x is a random variable in the sense of the above definition.

The probability $P\{a \leqq x < b\}$ considered as a function of a and b will be called the *integral probability law* of x.

A random variable is contrasted with a constant, say θ, the numerical values of which, corresponding to all elements of the set (A), are all equal. If θ is a constant, then whatever $a < b$ and B, the probability $P\{a \leqq \theta < b \mid B\}$ may have only values unity or zero according to whether θ falls in between a and b or not.

If we keep in mind the above definitions of the variables in our discussions of them, we may speak in terms of random experiments. In the sense of the convention adopted previously, we may say that x is a random variable when its values are determined by the results of a random experiment.

It is important to keep a clear distinction between *random variables* and *unknown constants*. The 1000th decimal, X_{1000}, in the expansion of $\pi = 3.14159 \cdots$ is a quantity unknown to me, but it is not a random variable since its value is perfectly fixed, whatever fundamental probability set we choose to consider. We could say alternatively that the value of X_{1000} does not depend upon the result of any random experiment.

Frequently we have to consider simultaneously several random variables

$$x_1, x_2, \cdots, x_n \tag{18}$$

and their simultaneous integral probability law, to be defined as follows.

Denote by E the set of values of the n variables (18). This set could be represented by a point which will be called the sample point E in an n-dimensional space, say W, the rectangular coordinates of the point E being the values $x_1, x_2, \cdots, x_n$. The space W will be called the sample space. Denote by w any region in W and accept the convention that $E \epsilon w$ stands for the words: "the point E is an element of w."

If the x_i are random variables, then whatever be w, we may speak of the probability of E being an element of w, and denote it by $P\{E \epsilon w\}$. In fact this probability will be represented by the ratio of the measure of that part, say $F(w)$, of the F.P.S. in which the x_i have values locating the point E within the boundaries of w to the measure of the F.P.S. itself. Of course, it must be assumed that w is measurable. With that restriction the probability, $P\{E \epsilon w\}$, is defined for every region w. This probability, considered as a function of the region w, is called the *simultaneous integral probability law* of the x_i.

Apart from, or instead of, the integral probability law we may frequently consider another function called the *elementary probability law* of the random variables. This is defined as follows.

If $P\{E \epsilon w\}$ stands for the integral probability law of the variables (18), and if there exists a function $p(E)$ of the x_i such that whatever be w, for which the probability $P\{E \epsilon w\}$ exists,

$$P\{E \epsilon w\} = \iint \cdots \int_w p(E)\, dx_1, dx_2 \cdots dx_n, \tag{19}$$

then the function $p(E)$ is called the *elementary probability law* of the random variables (18).

Remark: The terms "integral probability law" and "elementary probability law" were introduced in the 1920's by the noted French probabalist, Paul Lévy. In more recent times they are being partially replaced by "distribution function" and "probability density function," respectively.

It will be noticed that while the integral probability law is a function of the region w, the elementary probability law is a function only of the point E. It will be noticed also that $p(E)$ may be considered as being defined in the whole sample space and non-negative. Of course there are cases where no elementary probability law in the above sense exists; this, however, happens rarely in problems of statistics.

It is important to know a few simple rules of dealing with elementary probability laws.

(i) If $p(x_1, x_2, \cdots, x_n)$ and $p(x_1, x_2, \cdots, x_{n-1})$ are the elementary probability laws of

$$\left.\begin{array}{r} x_1, x_2, \cdots, x_{n-1}, x_n \\ \text{and} \\ x_1, x_2, \cdots, x_{n-1} \end{array}\right\} \tag{20}$$

respectively, then

$$p(x_1, x_2, \cdots, x_{n-1}) = \int_{-\infty}^{\infty} p(x_1, x_2, \cdots, x_{n-1}, x_n)\, dx_n. \tag{21}$$

This rule permits the calculation of the elementary probability law of any single one of the x_i whenever their simultaneous probability law is known.

(ii) If there are two sets of n random variables each,

$$x_1, x_2, \cdots, x_n \tag{22}$$

and

$$y_1, y_2, \cdots, y_n \tag{23}$$

such that each of the x_i is a function of the y_i, possessing continuous partial derivatives with regard to any y_i, the Jacobian

$$\Delta = \frac{\delta(x_1, x_2 \cdots x_n)}{\delta(y_1, y_2 \cdots y_n)} \tag{24}$$

existing and being different from zero almost everywhere and never changing its sign, then the probability laws $p(x_1, \cdots, x_n)$ and $p(y_1, \cdots, y_n)$ of the variables (22) and (23) respectively, are connected by the identity

$$p(y_1, y_2, \cdots, y_n) = p(x_1, x_2, \cdots, x_n)\left|\, \Delta \,\right| \tag{25}$$

where in the right-hand side the x_i will ordinarily be expressed in terms of the y_i.

Combining the two above rules we may calculate the probability law of various functions, $f(E)$, of the x_i whenever the simultaneous probability law of the latter is known.

In order to clear the way for the material involved in the following lectures, I shall finish this one by giving definitions relating to statistical hypotheses.

Consider the set of random variables $x_1, x_2, \cdots, x_n$. Any assumption concerning their probability law (either integral or elementary) is called a statistical hypothesis.

A statistical hypothesis is called *simple* if it specifies the integral probability law, $P\{E \epsilon w\}$ of the x_i as a single-valued function of the region w.

Any statistical hypothesis that is not simple is called *composite*. It may be useful to illustrate these definitions by some examples.

The assumption H_1 that [13]

$$p_X(E) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\Sigma(x_i-\mu)^2/2\sigma^2}, \tag{26}$$

where neither μ nor $\sigma > 0$ is specified, is a composite statistical hypothesis. In fact, if w denotes a region defined by the inequality

$$\Sigma x_i^2 < 1,$$

then

$$P\{E \epsilon w\} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \int\cdots\int_w e^{-\Sigma(x_i-\mu)^2/2\sigma^2} dx_1\, dx_2 \cdots dx_n \tag{27}$$

is not uniquely determined but is a function of the parameters μ and σ, which are left unspecified by the hypothesis H_1.

On the other hand, the assumption H_2 that the elementary probability law of the x_i is as given by formula (26) but with $\mu = 0$ and $\sigma = 1$ is already a simple hypothesis. In fact, whatever the region w in the sample space, substituting $\mu = 0$ and $\sigma = 1$ in (27), we shall be able to calculate the unique numerical value of $P\{E \epsilon w\}$, although at times this may be connected with great technical difficulties.

Part 2. Probability and Experimentation

1. Abstract character of mathematical theories and possibilities of applications. It is probable that many who listened to my first lecture were disappointed. They are engaged in applying probability to practical problems and such problems may be the only cause of their interest in the

[13] The sign Σ, unless accompanied by other indications, will signify summation over i from 1 to n; i.e., $i = 1, 2, \cdots, n$.

theory of probability. They may feel that they have no use for a theory which treats "experiments," "results," or in fact everything that is of the utmost importance to them only as picturesque descriptions of probability sets and measures. Theory of this kind may be good for mathematicians, they may say, but we want a mathematical theory dealing with actual experiments, not with abstract probability sets.

It may be useful to start this lecture by considering more closely whether or not it is possible to satisfy that part of my audience which is of the opinion described. One might put the question this way: Is it possible to produce a mathematical theory dealing with actual experiments or, more generally, with phenomena of actual life?

My answer is: Probably never. That is, unless the word mathematics changes its present meaning. The objects in a real world, or rather our sensations connected with them, are always more or less vague and since the time of Kant it has been realized that no general statement concerning them is possible. The human mind grew tired of this vagueness and constructed a science from which anything that is vague is excluded—this is mathematics. But the gain in generality must be paid for, and the price is the abstractness of the concepts with which mathematics deals and the hypothetical character of the results: *if* A is B and B is C, *then* A is also C.

Of course, there are many mathematical theories that are successfully applied to practical problems. But this does not mean that these theories deal with real objects. If they did, they could not involve general statements and could not be considered as mathematical. Let us illustrate this by a few examples. Modern geometry is a mathematical science and is applied to practical problems. But does it deal with objects that we meet in actual life? Let us see. Geometry deals with such concepts as planes, straight lines, points, etc. Is there anything in real life that is exactly a plane in the sense of geometry? We say sometimes that the surface of a table is a plane. But if we look at the surface through a good magnifying glass we shall immediately see that it is certainly not a plane. If we say that it is, we mean that *for practical purposes* it could be considered a plane.

Here we come to the essential point: when we apply mathematics to practical problems we never seek (and if we would, we should never succeed) to find an identity between mathematical concepts and realities; we are satisfied if we find some correspondence between them, by which a mathematical formula can be interpreted in terms of realities and give a result which, *for practical purposes,* would in our opinion be sufficiently accurate.

Consider a triangle T_1 formed by three points on this sheet of paper. Divide it by straight lines into four smaller triangles T_2, T_3, T_4 and T_5. If we state numerically the coordinates of all the vertices, we shall be able

to apply known formulas and calculate the areas of all the five triangles. Naturally, the area of T_1 will be equal to the sum of the areas of the other four. This is geometry. But now take any instruments you desire and measure the sides of all the triangles as actually drawn. Using these measurements and again applying formulas we may be disappointed to find that the area of T_1 so calculated is not exactly equal to the sum of the areas of T_2, T_3, T_4 and T_5.

It will be suggested that the discrepancy is due to errors of measurement. This is true so far as the expression "errors of measurements" stands for something broader, including the fact that the dots representing the vertices of the triangles are not the points we consider in mathematics. However, *for many practical purposes* the agreement between the area of T_1 and the sum of areas of T_2, T_3, T_4 and T_5 will be judged satisfactory and this is the decisive point in the question of whether or not the mathematical theory of geometry can be applied in practice.

A closer examination of other mathematical theories applied to practical problems will reveal the same features. The theory itself deals with abstract concepts not existing in the real world. But there are real objects that correspond to these abstract concepts in a certain sense, and numerical values of mathematical formulas more or less agree with the results of actual measurements. In the earlier stages of any branch of mathematically treated natural science we are satisfied with only a slight resemblance between mathematical and empirical results, but later on our requirements become more and more stringent.

After this somewhat long general introduction we may turn to the main topic of this lecture which is whether, and if so, how the mathematical theory of probability can be usefully applied in natural science.

2. Random experiments and the empirical law of large numbers. It follows from what I said that the foundations of the theory of probability could be chosen in many ways. But however they are chosen, if their accuracy is on the level now customary in mathematics, the theory of probability will deal with abstract concepts and not with any real objects. Therefore, the application of such a theory will be possible only if one can establish a bridge or a correspondence between concepts of the theory and real facts. The actual applications must be preceded by numerous checks and rechecks of the permanency and the accuracy of such correspondence. If one judges this to be sufficiently accurate and finds it sufficiently permanent, then the predictions—the final aim of any science—based on the mathematical theory of probability, will have some prospect of success. Otherwise the theory may be interesting by itself, but useless from the point of view of application.

What, then, is the class of facts that corresponds to concepts of the

theory of probability as described in my first lecture? What is the meaning of this correspondence?

The class of such facts may be described as the results of random experiments. It is impossible to give an exact definition of experiments that are called random, but it would be equally impossible to give a definition of objects in the real world that deserve the description "plane," "straight line," etc. If we try to do so, we shall inevitably find ourselves speaking not of real objects but of abstract concepts. At most we can give a rough description of random experiments and some illustrations so as to appeal to the intuition. In what follows, unless otherwise stated, whenever I shall speak of experiments I shall mean real experiments, not hypothetical ones.

There are experiments which, even if carried out repeatedly with the utmost care to keep conditions constant, yield varying results. They are "random."

(a) We may construct a special machine to toss coins. This machine may be very strong, driven by an electric motor so as to impart a constant initial velocity to the coin. The experiments may be carried on in a closed room with no noticeable air currents; the coin may be put into the machine always in the same way; and even then I am practically certain the results of the repeated experiments will vary. Perhaps frequently we may get heads, but from time to time the coin will fall tails. The experimenter may be inclined to think that these cases arise from some "error of experimentation."

(b) Another example of this kind is provided by roulette. A well-constructed roulette wheel with an electrically regulated start will yield varying results.

(c) The above were types of random experiments arranged by men. But there are some going spontaneously. Consider a quantity of radioactive matter and the α particles it emits in some specified direction within a cone of small solid angle. These particles could be recorded by the fluorescence they produce when falling on an appropriate screen. Let us observe this screen for several consecutive minutes, one minute's observation being considered as a single experiment. It will be found that however constant be the conditions of the consecutive experiments, the results will vary in that the number of disintegrations recorded per minute will not be the same.

(d) Another example of this kind is provided by the varying properties of organisms forming an F_2 generation, however homogeneous be the conditions of breeding.

These examples may make sufficiently clear what I mean by random experiments. Now I shall explain the sense in which their results correspond to concepts involved in the theory of probability.

Let N and n be positive integers, N fairly large, say 1000 or so, and n moderate, say 10. Let us perform a long series of Nn random experiments of the type described, and count cases where a certain specified result E occurred. Let it be in M cases. Dividing M by Nn we obtain the ratio

$$f = \frac{M}{Nn} \tag{1}$$

which will be called the relative frequency of the result E in the course of Nn trials. These Nn trials will be called experiments of the first order. Now divide the whole series of Nn first order experiments into N groups of n trials each in the order in which the trials were carried out. Each such group of n first order trials will now be considered as a trial of second order.

The second order trials could be classified according to the number k of occurrences of the result E in the n first order trials of which they are formed. Obviously k could be equal to 0, 1, 2, $\cdots$, n, in any one of the second order trials. Let m_k denote the number of trials in which E occurred exactly k times, and

$$F_{n,k} = \frac{m_k}{N} \tag{2}$$

the relative frequency in the series of second order trials.

It is a surprising and very important empirical fact that whenever sufficient care is taken to carry out the first order experiments under as uniform conditions as possible, and when the number N is large, then the relative frequency $F_{n,k}$ appears to be very nearly equal to the familiar formula

$$\frac{n!}{(n-k)!k!}(1-f)^{n-k}f^k. \tag{3}$$

In other words, the relative frequency $F_{n,k}$ relating to a series of second order experiments is connected with the relative frequency of the first order experiments in very nearly the same way as the probability $P_{n,k}$ relating to the second order probability set, as discussed in my first lecture, is connected with the probability p referring to the corresponding first order probability set.

In order to avoid misunderstanding, let us describe the situation in greater detail. Suppose that the random experiment under consideration consists in $2N$ throws of the same die and that f is the relative frequency of cases where the upper side of the die had six points on it. The value of f may be close to 1/6 or not. It may, in fact, differ considerably from 1/6, depending on the structure of the die and the exact conditions of throwing. But if we split the whole series of trials into consecutive pairs, then the proportions of pairs with 0, 1 and 2 sixes will be, approximately,

$$(1-f)^2, \quad 2f(1-f) \quad \text{and} \quad f^2. \tag{4}$$

The above fact, which has been found empirically [1] many times, could be described in a more general way by saying that single random experiments and the various groups of these experiments usually behave as if they tended to reproduce certain first order probability sets, corresponding to first order trials, and the appropriate second order probability set. This fact may be called the empirical law of large numbers. I want to emphasize that this law applies not only to the simple case connected with the binomial formula which was discussed above but also to other cases. In fact, this law seems to be perfectly general, in the sense in which we use the word general with respect to any other "general law" observed in the outside world. Whenever the law fails, we explain the failure by suspecting a "lack of randomness" in the first order trials.

Suppose now that having repeatedly performed series of random experiments of some specified kind we have always found that they do conform to the empirical law of large numbers. Then, as is our custom, we expect them to behave similarly in the future, and we expect the calculus of probability to permit us to make successful predictions of frequencies of results of future series of experiments.

This is the way in which the abstract theory of probability described in my first lecture may be put into correspondence with happenings in the outside world and how it may be, and actually is, applied to solve problems of practical importance. The standing of the theory of probability is, in this respect, no different from any other branch of mathematics. The application of the theory involves the following steps.

(i) If we wish to treat certain phenomena by means of the theory of probability we must find some element of these phenomena that could be considered as random, following the law of large numbers. This involves a construction of a mathematical model of the phenomena involving one or more probability sets.

(ii) The mathematical model may be satisfactory or not. This must be checked by observation.

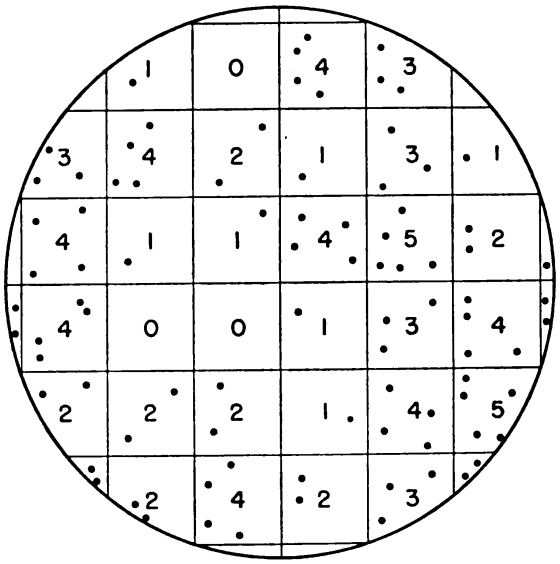
(iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future.

Let us illustrate these steps by a few examples taken from the current literature.

3. Illustrations. *Example 1.*—Two bacteriologist friends of mine, Miss J. Supinska and Dr. T. Matuszewski, were interested in learning whether the calculus of probability could be applied to certain problems concerning

[1] See, for example, L. von Bortkiewicz, *Die Iterationen,* Julius Springer, Berlin, 1917, x + 205 pp.

the colonies of bacteria on a Petri-plate. The diagram reproduces a photograph of a Petri-plate with colonies that are visible as dark spots.



You will notice that the plate is divided into a number of small squares. In order to explain the particular mathematical model that was tried in this instance, consider the contents v of one particular square and consider one particular living bacterium B contained in the liquid that was poured on the plate. In the mathematical model all the operations performed with the liquid and the plate which resulted in fixing the bacterium B in some point are considered as a first order experiment which may result either in B falling within v, or not. If there were N living bacteria in the liquid poured on to the plate, then there were N such first order experiments all relating to the same square v. They form a single second order experiment. Finally, if the number of squares in which the plate is divided be n, then there will be n second order experiments, which, taken together, could be considered as one third order experiment. Without going into further details of this mathematical model I shall state that it implies that the probability of any of the squares containing exactly k colonies must be approximately equal to the Poisson formula

$$P_k = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{5}$$

where λ means the average number of colonies per square. The reader will notice that the above k satisfies the definition of a random variable the integral probability law of which is given by

$$P\{a \leqq k < b\} = \sum_{k=a}^{b-1} e^{-\lambda} \frac{\lambda^k}{k!} \qquad \text{for } 0 \leqq a < b. \tag{6}$$

If this mathematical model could be assumed to correspond accurately to the actual experiments in the sense explained above, then it could be used for predicting frequencies of certain circumstances that are important in bacteriology. One of the questions that my colleagues had in mind was how frequently a single colony is produced by two or more unconnected bacteria.

In order to answer the question whether or not the number k of colonies within a square could be considered as a random variable whose probability law could be represented by formula (5), my colleagues performed a series of experiments summarized in Table I.

The values of k are the numbers of colonies within the squares into which the whole plate was divided. m' and m denote respectively the observed and the expected numbers of squares having the number k of colonies. The last two lines give measures of the goodness of fit, the chi-square and the corresponding P. It is seen that without exception the agreement between the observed and the theoretical frequencies obtained by multiplying the P_k of formula (5) by the total number of squares on the plate, is surprisingly good. As a matter of fact, the total number of similar experiments that have been carried out is much larger, and in not a single case has any serious disagreement between the distribution of colonies and the Poisson law been recorded. This entitles us to expect that the results of future experiments will be similar, and that conclusions concerning these future experiments drawn from the mathematical model described above, will be correct, or good enough.

If the model implies that in a particular case the probability of a colony arising from more than one independently floating individual is for instance $P = .001$, we may conclude that about 99.9 percent of the colonies were produced by one individual only.

For the sake of clearness I may mention that in the above statement "one individual" does not necessarily mean one cell. This expression refers to one or more cells that are floating together, being connected either mechanically or biologically.

Example 2.—Table II is reproduced from an article in *Biometrika,* and represents a comparison between the Poisson law, formula (5), and the distribution of dodder in samples of clover seed. The problem and the mathematical model were similar to that treated above.

The table gives 12 comparisons, of which eleven are based on material produced by Schindler and the last by the authors of the article, J. Przyborowski and H. Wilenski. It will be seen that the material as a whole is

TABLE I

Comparison of distribution of colonies with Poisson Law

[T. Matuszewski, J. Supinska and J. Neyman, *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten.* II. Abteilung, 1936, Bd. 95].

k	Plate 1		Plate 2		Plate 3		Plate 4		Plate 5	
	m'	m	m'	m	m'	m	m'	m	m'	m
0	5	6.1	26	27.5	59	55.6	83	75.0	0	0.7
1	19	18.0	40	42.2	86	82.2	134	144.5	5	3.9
2	26	26.7	38	32.5	49	60.8	135	139.4	9	11.0
3	26	26.4	17	16.7	30	30.0	101	89.7	23	20.9
4	21	19.6	5		15		40	43.3	33	29.6
5	13	11.7	2	+9.1	3	+15.4	16	16.7	32	34.0
6	4				2		3		32	31.8
7	3	+9.5					2	+7.4	24	25.8
8	1						2		13	18.3
9									12	11.6
10									8	6.7
11									7	
12									2	+5.7
χ^2		0.77		1.61		4.05		3.47		4.94
P_{χ^2}		0.97		0.66		0.26		0.63		0.84

k	Plate 6		Plate 7		Plate 8		Plate 9		Plate 10	
	m'	m	m'	m	m'	m	m'	m	m'	m
0	8	6.8	0	10.3	7	3.9	3	2.1	60	62.6
1	16	16.2	12		11	10.4	7	8.2	80	75.8
2	18	19.2	18	16.7	11	13.7	14	15.8	45	45.8
3	15	15.1	13	22.4	11	12.0	21	20.2	16	18.5
4	9	9.0	27	22.7	7	7.9	20	19.5	8	
5	4		19	18.3	3		19	15.0	1	+7.3
6	2		16	12.3	2		7	9.6		
7	0	+6.7	6		1	+7.1	6			
8	1		4	+13.3	1		1			
9			1		1		0	+9.6		
10							2			
χ^2		0.30		6.67		3.21		2.63		1.09
P_{χ^2}		0.97		0.25		0.53		0.85		0.78

k = number of colonies per square.
m' = observed frequency.
m = expected frequency (Poisson).

TABLE II

Comparison of the distribution of dodder seeds in samples of clover with Poisson Law *

[J. Przyborowski and H. Wilenski, *Biometrika*, Vol. 27, 1935, p. 277]

k	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$
0	168	183.94	599	606.53	382	389.40	284	303.27	795	774.64
1	205	183.94	315	303.27	111	97.35	170	151.63	94	116.20
2	94	91.97	74	75.82	7	12.17	39	37.91	11	8.71
3	26	30.66	12	12.64		+1.08	7	6.32		0.45
4	6	7.66		+1.74				+0.87		
5	1	1.53								
Over 5		+0.30								
χ^2		5.20		0.98		5.00		3.49		5.13
P_{χ^2}		0.160		0.600		0.000		0.180		0.000

k	Sample 6		Sample 7		Sample 8		Sample 9		Sample 10	
	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$	N_k	$N \cdot P_k$
0	447	452.42	473	475.61	295	303.27	22	16.42	0	1.08
1	51	45.24	26	23.78	153	151.63	29	41.04	3	5.39
2	2	2.26	+1	+0.61	44	37.91	55	51.30	13	13.48
3		+0.08			8	6.32	43	42.75	15	22.46
4						+0.87	34	26.72	33	28.07
5							10	13.36	28	28.07
6							3	5.57	24	23.40
7							4	1.99	21	16.71
8								0.85	10	10.44
9									8	5.80
10									+5	+5.10
χ^2		0.85		0.47		1.31		8.76		7.04
P_{χ^2}		0.198		0.319		0.533		0.120		0.532

k = number of dodder seeds in a sample.
N_k = observed frequency.
$N \cdot P_k$ = expected frequency (Poisson).
* Data for the first eleven samples are taken from Schindler's experiments.

TABLE II—Continued

k	Sample 11		Authors' own experiment with known $\lambda = 2$		
	N_k	$N \cdot P_k$	k	N_k	$N \cdot P_k$
0	0	0.09	0	56	67.67
1	0	0.66	1	156	135.34
2	1	2.49	2	132	135.34
3	4	6.22	3	92	90.22
4	9	11.67	4	37	45.11
5	16	17.50	5	22	18.04
6	19	21.87	6	4	6.02
7	19	23.44	7	0	1.72
8	26	21.97	8	1	0.43
9	19	18.31	Over 8	0	0.12
10	15	13.73			
11	14	9.36			
12	5	5.85			
13	6	3.38			
14	3	1.81			
15	3	0.90			
Over 15	+1	+0.74			
χ^2		9.81			8.92
P_{χ^2}		0.548			0.179

not as satisfactory as in the preceding example. It seems to follow that if samples of clover seed are drawn by the method employed by Schindler, then conclusions concerning them drawn from the mathematical model involving the Poisson Law will not necessarily be very accurate. But it is possible that the method of drawing samples of seeds may be so adjusted (this is the opinion of Przyborowski and Wilenski) that the number of dodders in a small subsample of seeds could be considered rightly as a random variable following the Poisson Law.

As mentioned above, if the outcomes of experiments or observations do not conform with the predictions of a mathematical model that is strongly suggested by intuition, then it is usual to ascribe the divergencies to "faults of experimentation." This expression is vague, and if we try to make it more precise, we shall probably come to the description: "The random machinery of the observed phenomena does not correspond to the mathematical model assumed." The situation can be remedied in two ways. One is to make an effort towards a better understanding of the phenomena

studied and therewith to modify the mathematical model. The other way is to modify the method of experimentation so as to bring it into conformity with the original mathematical model. The possibility and desirability of these two methods depend on the circumstances of the problem. They are illustrated in the following two examples.

Example 3.—Problems of pest control led to studies of the distribution of larvae in small plots. An experimental field planted with some crop is divided into a number of small plots, very much as a Petri-plate in Example 1 was divided into small squares. Then all the larvae found in each plot are counted. Naturally, the number of larvae varies considerably from one plot to another. The original mathematical model of the machinery behind this variability, the one strongly suggested by intuition, was the same as that used for the interpretation of the variability of the number of colonies from one square on the Petri-plate to another. Therefore, attempts were made to fit the observed distributions with a Poisson frequency law. Counts of larvae and attempts to understand the machinery of their distributions were made by many research workers. Table III,

TABLE III

Comparison of the distribution of beet web worms with the Poisson and Type A contagious distributions

[G. Beall, *Ecology*, Vol. 21, 1940, p. 462]

Class	Treatment 1 (untreated)			Treatment 2			Treatment 3		
	Obs.	Poisson exp.	Type A exp.	Obs.	Poisson exp.	Type A exp.	Obs.	Poisson exp.	Type A exp.
0	117	80.1	116.7	205	196.2	203.8	162	138.6	157.6
1	87	112.2	84.3	84	99.0	87.8	88	118.1	96.0
2	50	78.5	58.3	30	25.0	25.9	45	50.3	45.4
3	38	36.7	33.6	4	4.2	6.1	23	14.3	17.6
4	21	12.8	17.4	2	0.5	1.2	5	3.0	6.0
5	7	3.6	8.3				2	0.5	
6	2	0.8	3.7		+0.1	+0.2			
7	2	0.2	1.6					+0.2	+2.4
8	0								
9	1	+0.1	+1.1						
m_1			2.114			3.204			2.537
m_2			0.662			0.157			0.336
χ^2		46.8	3.1		4.0	1.1		20.2	2.7
P_{χ^2}		0.000	0.543		0.135	0.282		0.000	0.269

TABLE III—Continued

Comparison of the distribution of diplopods with the Poisson and Type A contagious distributions

[L. C. Cole, *Ecological Monographs*, Vol. 16, 1946, p. 71]

Number per board	Obs.	Poisson exp.	Type A exp.
0	128	100.5	133.6
1	71	95.5	61.0
2	34	45.4	35.6
3	11	14.4	17.2
4	8	3.4	7.5
5	5	0.7	3.1
Over 5	3	0.1	2.0
m_1			1.307
m_2			0.712
χ^2		20.5	4.1
P_{χ^2}		0.000	0.249

taken from data in papers by Geoffrey Beall,[2] Lamont C. Cole [3] and S. B. Fracker and H. A. Brischle,[4] gives a few observed distributions and their comparison with theoretical distributions.

In all cases, the first theoretical distribution tried was that of Poisson. It will be seen that the general character of the observed distribution is entirely different from that of Poisson. There seems to be no doubt but that a very serious divergence exists between the actual phenomenon of distribution of larvae and the machinery assumed in the mathematical model. When this circumstance was brought to my attention by Dr. Beall, we set out to discover the reasons for the divergence.

From the discussion of Example 1 you will perceive that, if we attempt to treat the distribution of larvae from the point of view of Poisson, we would have to assume that each larva is placed on the field independently of the others. This basic assumption was flatly contradicted by the life of larvae as described by Dr. Beall. Larvae develop from eggs laid by

[2] Geoffrey Beall: "The fit and significance of contagious distributions when applied to observations on larval insects." *Ecology*, Vol. 21 (1940), pp. 460–474.

[3] Lamont C. Cole: "A study of cryptozoa of an Illinois woodland." *Ecological Monographs*, Vol. 16 (1946), pp. 49–86.

[4] S. B. Fracker and H. A. Brischle: "Measuring the local distribution of Ribes." *Ecology*, Vol. 25 (1944), pp. 283–303.

TABLE III—Continued

Comparison of distribution of ribes with the Poisson and Type A contagious distributions

[S. B. Fracker and H. A. Brischle, *Ecology*, Vol. 25, 1944, p. 291]

Number per 0.1 acre strip	Obs.	Poisson exp.	Type A * exp.
0	42	18.9	42.3
1	11	23.0	15.6
2	4	14.0	10.8
3	1	5.7	5.9
4	3	1.7	3.0
5	1	0.4	1.4
6	0	0.1	
Over 6	2	+0.2	+1.0
m_1			1.000
m_2			1.013
χ^2		41.7	1.92
P_{χ^2}		0.000	0.392

* In the original publication the fit given was worse, due to maladjustment of parameters m_1 and m_2.

moths. It is plausible to assume that, when a moth feels like laying eggs, it does not make any special choice between sections of a field planted with the same crop and reasonably uniform in other respects. Therefore, as far as the spots where a number of moths lay their eggs is concerned, it is plausible that the distribution of spots follows a Poisson Law of frequency, depending on just one parameter, say m, representing the average number of spots per unit area.

However, it appears that the moths do not lay eggs one at a time. In fact, at each "sitting" a moth lays a whole batch of eggs and the number of eggs varies from one cluster to another. Moreover, by the time the counts are made the number of larvae is subject to another source of variation, due to mortality.

After hatching in a particular spot, the larvae begin to look for food and crawl around. Since the speed of their movements is only moderate, it is obvious that for a larva to be found within a plot, the birthplace of this larva must be fairly close to this plot. If one larva is found, then it is likely that the plot will contain more than one from the same cluster.

Considerations of this kind were used to build up a mathematical model of the distribution of larvae which led to the following results. Let $C(k)$ denote the probability that a plot will contain exactly k larvae, for $k = 0, 1, 2, \cdots$. The probability $C(0)$ that there will be no larvae in the plot considered is computed from the formula

$$C(0) = e^{-m_1(1-e^{-m_2})}. \tag{7}$$

If $C(0), C(1), \cdots, C(k)$ are computed, then $C(k+1)$ is given by the recurrence formula

$$C(k+1) = \frac{m_1 m_2 e^{-m_2}}{k+1} \sum_{t=0}^{k} \frac{{m_2}^t}{t!} C(k-t). \tag{8}$$

In particular,

$$C(1) = e^{-m_1(1-e^{-m_2})} \frac{m_2}{1!} m_1 e^{-m_2}, \tag{9}$$

$$C(2) = e^{-m_1(1-e^{-m_2})} \frac{{m_2}^2}{2!} ({m_1}^2 e^{-2m_2} + m_1 e^{-m_2}), \tag{10}$$

etc.

It may be regretted that the formulae are somewhat complicated. However, since the machinery behind the distribution of larvae is rather complex, one has to put up with the resulting inconvenience.

Because, as we have observed, a plot that contains one larva frequently contains more than one, the distribution deduced was called "contagious." Several distributions of a similar kind were deduced and, to make a distinction, the one given by the above formulae was called contagious of type A with two parameters.[5]

A distribution of type A depends on two parameters, m_1 and m_2, which are connected with three quantities having a physical meaning as follows. Assume that the area of the plot on which the larvae are counted is equal to unity. Further, let m be the average number of batches of eggs per unit of area, and let λ be the average number of survivors per batch of eggs at the time when the counts are made. Finally, let us introduce an area A which we shall call "area of accessibility." Imagine a plot P of unit area on which counts of larvae are to be made and let S denote a spot on which a batch of eggs was laid. If S is far from P, then no larva hatched at S can be found in P. The area A, by definition, contains all points S such that larvae born at S *can* reach the plot P before the counts are made.

[5] The term "contagious distribution" was borrowed from G. Pólya, who was the first to consider this type of problem. See G. Pólya: "Sur quelques points de la théorie des probabilités." *Annales de l'Institut Henri Poincaré,* Vol. 1 (1931), pp. 117–162.

See also W. Feller: "On a general class of 'contagious' distributions." *Annals of Math. Stat.,* Vol. 14 (1943), pp. 389–400.

Obviously, the more mobile the larvae are, the larger the area A and conversely. Consequently, if one counts very young larvae, then A is small, close to unity. For larger larvae, the area A is larger. It follows that a reasonable agreement between theory and observation may be expected only if counts include larvae of more or less the same age.

The parameters m_1 and m_2 are connected with m, λ, and A by the following formulae:

$$m_1 = Am, \qquad m_2 = \frac{\lambda}{A}. \tag{11}$$

The mean number of larvae per plot is

$$\mu_1' = \lambda m = m_1 m_2, \tag{12}$$

the variance is

$$\mu_2 = \lambda m\left(1 + \frac{\lambda}{A}\right) = m_1 m_2(1 + m_2). \tag{13}$$

It is seen that if the mean μ_1' is kept constant while the area of accessibility A is indefinitely increased, then the contagious distribution approaches the Poisson Law. Details concerning the distribution can be found in the original publication.[6] Table III gives the comparison between the observed distribution of larvae and the one expected on the basis of contagious distribution of type A with two parameters. It is seen that in all cases the agreement is satisfactory. The data presented do not exhaust the instances where contagious distributions of type A fit actual counts of insects. In fact, it seems already safe to say that satisfactory agreement between this particular mathematical model and observation is a more or less general rule with the restriction that the life of the insects concerned does not depart too widely from the general scheme described above. On the other hand, there are organisms (e.g., scales) whose distribution on units of area of their habitat does not conform with type A. An investigation revealed that the processes governing the distribution of these organisms were much more complex than that described and therefore, if a statistical treatment is desired, a fresh effort to construct an appropriate mathematical model is necessary.

In this example, in order to have agreement between the observed and predicted frequencies, it was imperative to adjust the mathematical model. This is generally the case when the phenomena studied develop by themselves and do not admit of any sort of human control. In the next example we consider an instance of another kind where the experimental technique may be so changed as to fit a desirable mathematical model.

[6] J. Neyman: "On a new class of 'contagious' distributions, applicable in entomology and bacteriology." *Annals of Math. Stat.*, Vol. 10 (1939), pp. 35–57.

Example 4.—This example deals with a category of industrial problems. Problems of this kind are treated by Walter A. Shewhart [7] and the reader will find them of considerable interest.

Many laboratories are engaged in what is called routine analysis. Small quantities of certain materials are sent to the laboratory for determining the content of a certain ingredient X. The sample is subdivided into a few portions, three, four or sometimes five, and these are analyzed separately. Denote the particular results by x_1, x_2, x_3 and x_4 respectively and by μ the "true" content of the ingredient X so that the x_i denote the measurements of μ.

Because of experimental errors the measurements x_i differ from μ and differ among themselves. Frequently there is evidence that the measurements could be regarded as random variables following a normal law of frequency,

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \tag{14}$$

so that this formula forms the mathematical model of the experiments of first order. The model may be used to estimate the value of μ knowing only the values of four measurements x_1, x_2, x_3 and x_4. But we can proceed differently. Denote by f_1 and f_2 some two functions of the x_i. If the x_i are random variables, then f_1 and f_2 will also be random variables and we may consider probabilities of their satisfying any given inequalities. We may also look for some particular forms of the functions f_1 and f_2 such that the probability of their satisfying a given inequality shall be equal to any given number between zero and unity. Starting from this point of view it has been found that the functions [8]

$$f_1 = \bar{x} - \frac{t_\alpha s}{\sqrt{n}}$$

and (15)

$$f_2 = \bar{x} + \frac{t_\alpha s}{\sqrt{n}}$$

have a remarkable property. Here $\bar{x}$ is the arithmetic mean of the measurements x_i, n their number, s their estimated standard deviation,[9] and t_α the

[7] Walter A. Shewhart: *The Economic Control of Quality of Manufactured Product.* Van Nostrand, New York, 1931, 501 pp.

[8] J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability." *Phil. Trans. Royal Soc.*, A236 (1937), pp. 333–380. See also the conferences on estimation and confidence intervals.

[9] That is, s is an estimate of σ; $s^2 = \Sigma\frac{(x_i - \bar{x})^2}{(n-1)}$.

value of Fisher's t corresponding to the number of degrees of freedom on which s is based, and to $P = 1 - \alpha =$, e.g., .01. If the measurements x_i are independent random variables following the normal law (14), then whatever be the values of μ and σ, the probability of f_1 falling short of μ and of f_2 exceeding μ is exactly equal to $\alpha = .99$.

This circumstance permits the estimation of μ in the form of a random experiment. We perform the experimental analysis, obtaining the values of the x_i, and then state that

$$\bar{x} - \frac{t_\alpha s}{\sqrt{n}} \leqq \mu \leqq \bar{x} + \frac{t_\alpha s}{\sqrt{n}}. \tag{16}$$

We may be wrong in this statement, but if the x_i do follow law (14), the probability of our being correct is equal to $\alpha = .99$. In other words, in 99 percent of such experiments, our statement concerning μ will be correct.

The arbitrarily chosen number α is called the *confidence coefficient* and the interval between f_1 and f_2 the *confidence interval.* If the number of measurements is small, something like $n = 4$, then the value of t_α is considerable, and the accuracy of estimating μ as measured by the length of the confidence interval

$$f_2 - f_1 = \frac{2t_\alpha s}{n} \tag{17}$$

is not satisfactory.

In what preceded, the value of σ in Equation (14) was considered unknown. If, however, σ is known, then the confidence interval will be written as

$$\bar{x} - \frac{T_\alpha \sigma}{\sqrt{n}} \leqq \mu \leqq \bar{x} + \frac{T_\alpha \sigma}{\sqrt{n}}, \tag{18}$$

where T_α is the value of t_α corresponding to an infinite number of degrees of freedom in the estimate of σ. What this means in practice may be judged from the following comparison. If $\alpha = .99$, then $T_\alpha = 2.576$, no matter what n is. At the same time the values of t_α are, respectively,

$$\left.\begin{aligned} t_{.01} &= 63.657 \quad \text{if } n = 2, \\ t_{.01} &= 9.925 \quad \text{if } n = 3, \\ t_{.01} &= 5.841 \quad \text{if } n = 4, \\ &\text{etc.} \end{aligned}\right\} \tag{19}$$

It follows that, whenever it is known not only that the analyses made in some particular laboratory provide numbers x that for practical purposes could be considered as particular values of a random variable following the normal probability law (14), but also that the standard deviation

σ has permanently some particular numerical value, then the same few parallel analyses could be used to provide an equally reliable but a much more accurate statement concerning the value of μ. Therefore, if a laboratory is permanently engaged in performing analyses of some particular kind, obviously it must be interested (i) in keeping the value of σ constant over long periods of time; (ii) in estimating this value of σ as accurately as possible; and (iii) in keeping watch over possible changes in σ.

In order to keep σ constant, say throughout a year, it is necessary to eliminate all factors that may influence the accuracy of the analyses. This is frequently done; but before trying to estimate the value of σ presumed to be constant, and before applying formula (18) instead of (16) we must see whether or not the measurements that are being obtained do agree with the mathematical model involving a constant σ. Otherwise, repeated application of formula (18) may give a much greater percentage of errors than that expected.

This circumstance was realized by J. Przyborowski, who published the following table illustrating his efforts to stabilize the accuracy of his analyses of oats. In Table IV, s_i^2 is the estimated variance of four parallel analyses, and s_0^2 is the arithmetic mean of a number of such variances calculated for a long period of time, such as a year or more. If the value of σ^2 were actually constant during such a period, then the value of s_0^2 would be a very accurate estimate and the mathematical model adopted would imply a known distribution of the ratio $v = s_i^2/s_0^2$.

The comparison of the expected and observed frequencies of the values of v are given in the table for various periods. And here we see the curious results of efforts to stabilize the accuracy of analyses. Year 1925 is very bad; 1927 and 1928 show slight improvement, but are still bad. 1929 and 1930 are excellent; but this probably caused a false sense of security of the personnel, and the next year 1931 is again bad. However, the three year period 1929–1931 seems to be satisfactory. We may reasonably hope that the experience of 1931 has stimulated the staff of Professor Przyborowski's laboratory and that confidence intervals based on formula (18), where the value of σ is estimated from a great number of previous experiments, do give correct statements concerning μ in nearly the expected percentage of cases, 100α.

4. Summary. Now let us sum up the main points that I have tried to emphasize. In speaking about probability, it is necessary to distinguish [10] three different but related aspects of the problem:

(1) a mathematical theory, for example, the one described in my first lecture;

[10] Compare with H. Levy and L. Roth, *Elements of Probability*. Clarendon Press, Oxford, 1936, p. 15.

TABLE IV

Distribution of estimated error variance in routine analyses of four parallel samples of oats

[J. Przyborowski, *Polish Agric. Forest. Journ.*, Vol. 30, 1933]

1925			1927			1928			1929			1930			1931			1929–1931		
$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.	$v = \frac{s_i^2}{s_0^2}$	Obs.	Exp.
0–1	76	35.6	0–1	44	24.8	0–1	49	31.0	0–1	35	30.8	0–1	25	27.0	0–1	20	20.7	0–1	80	78.5
1–2	30	41.0	1–2	27	28.6	1–2	31	35.7	1–2	27	35.5	1–2	33	31.1	1–2	20	23.8	1–2	81	90.4
2–3	20	32.4	2–3	13	22.6	2–3	27	28.2	2–3	30	28.0	2–3	27	24.6	2–3	15	18.8	2–3	71	71.4
3–4	14	23.3	3–4	19	16.3	3–4	18	20.3	3–4	19	20.2	3–4	20	17.7	3–4	22	13.5	3–4	61	51.4
4–5	10	16.1	4–5	7	11.2	4–5	9	14.0	4–5	13	13.9	4–5	13	12.2	4–5	15	9.3	4–5	40	35.4
5–6	4	10.8	Above 5	15	21.5	5–6	5	9.4	5–6	13	9.3	Above 5	18	23.4	Above 5	12	17.9	5–6	24	23.8
Above 6	25	20.0				Above 6	17	17.4	Above 6	18	17.3							6–7	14	15.7
																		7–8	13	10.2
																		Above 8	11	18.2
$\chi^2 = 65.101$ $P_{\chi^2} = 0.00000$			$\chi^2 = 22.928$ $P_{\chi^2} = 0.000127$			$\chi^2 = 15.217$ $P_{\chi^2} = 0.0094$			$\chi^2 = 4.332$ $P_{\chi^2} = 0.36$			$\chi^2 = 2.084$ $P_{\chi^2} = 0.72$			$\chi^2 = 12.068$ $P_{\chi^2} = 0.017$			$\chi^2 = 7.157$ $P_{\chi^2} = 0.41$		

(2) the frequency of actual occurrences;
(3) the psychological expectation of the participant.

The mathematical theory need not be the one I described but, if it is mathematically accurate, it will have nothing to do with the outside world and, therefore, with either (2) or (3). This is for the good reason that an accurate mathematical theory implies accurate definitions and axioms and that in the outside world there are no objects that satisfy them except within limits "good enough for practical purposes."

The theory of probability may be constructed to provide models corresponding in some sense to certain phenomena of the outside world. And here we may distinguish a divergence: (i) Some authors try to provide mathematical models of what I called random experiments, the aspect falling under (2) above. The theory presented in my first lecture is one of the types which comes under this heading. The theory of Richard von Mises is another. (ii) In building a mathematical theory of probability we may aim at a model of the changes in the state of the human mind concerning certain statements that occur as a result of changing the amount of known facts. This view is exemplified by the theory built by Harold Jeffreys.[11] It will be noticed that the theory of probability of my first lecture has nothing to do with a "state of mind," although, if we find that the probability of a certain property is equal to 0.0001, for example, the state of our mind will undoubtedly be influenced by this finding.

As I have mentioned, any theory may be correct if the authors are sufficiently accurate in their deductions. However, it is my strong opinion that no mathematical theory refers exactly to happenings in the outside world and that any application requires a solid bridge over an abyss. The construction of such a bridge consists first, in explaining in what sense the mathematical model provided by the theory is expected to "correspond" to certain actual happenings and second, in checking empirically whether or not the correspondence is satisfactory.

The examples which I have given and many others which could easily be quoted indicate that, by taking care both in the constructing of a mathematical model and in the carrying out of the experiments, the bridge between the theory of probability sketched in this chapter and certain fields of application may be very solid.

[11] See Jeffreys' *Scientific Inference,* University Press, Cambridge (Eng.), 1931, 247 pp. Also numerous papers in the *Proceedings of the Royal Society* (Series A) and in the *Proceedings of the Cambridge Philosophical Society.*

Part 3. Tests of Statistical Hypotheses

1. The traditional procedure in testing statistical hypotheses. The present lecture should not be considered as a direct continuation of the preceding ones which were systematically connected. However, the concepts discussed in my first two lectures will be used freely and combined with a few new ones. Since it would be impossible to give all the necessary definitions here, I must assume them to be known.

The traditional procedure in testing statistical hypotheses is widely known but, as it is traditional, opinions concerning its exact nature vary. I shall describe here a version that seems to summarize the common phases in the history of several well known tests, such as the chi-test for goodness of fit, Student's z test and others.

If we had to test any specified (in the early stages, very vaguely specified) statistical hypothesis H concerning the random variables,

$$x_1, x_2, \cdots, x_n,$$

we used to choose some function T of the x's which, for certain reasons, seemed to be suitable as a test criterion. Pearson's chi-square and Student's z are instances of such criteria. The next step, and sometimes a difficult one, consisted in deducing the exact probability law $p(T \mid H)$ or an approximate one, at least, which the chosen criterion T would follow if the hypothesis H were true. The graphs of the probability laws considered usually represented curves with a single maximum at a certain point of the range, decreasing towards the ends. This suggested a classification of possible samples into two not very distinctly divided categories, "probable" and "improbable" samples. If a sample E led to a value of the criterion T for which the value of $p(T \mid H)$ was small compared with its maximum, then the sample E would be called improbable, or the hypothesis H improbable, and conversely. You will certainly remember instances where both very small and very large values of chi-square are supposed to suggest that something is wrong.

When an "improbable sample" was obtained, the usual way of reasoning was this: "Were the hypothesis H true, then the probability of getting a value of T as or more improbable than that actually observed would be (e.g.) $P = 0.00001$. It follows that if the hypothesis H be true, what we actually observed would be a miracle. We don't believe in miracles nowadays and therefore we do not believe in H being true."

The above procedure, or something like it, has been applied since the invention of the first systematically applied test, the Pearson chi-square of

1900, and has worked, on the whole, satisfactorily.[1] However, now that we have become sophisticated we desire to have a theory of tests. Above all, we want to know why we should use this or that particular function T of the x's as a criterion. Why should we test the goodness of fit by calculating

$$\chi^2 = \Sigma \frac{(m - m')^2}{m} \tag{1}$$

and not, say

$$\chi'^2 = \Sigma \frac{(m - m')^2}{m'} \tag{2}$$

or

$$\chi''^2 = \Sigma \frac{|m - m'|}{m} \tag{3}$$

or something else? What is the actual meaning of a statistical test? What is the principle of choosing between several tests suggested for the same hypothesis? It is the purpose of the present lecture to discuss some of these questions and to explain certain basic ideas underlying the contributions to the theory of testing statistical hypotheses for which Professor E. S. Pearson and myself are responsible.

The first question I shall discuss is this: when selecting a criterion to test a particular hypothesis H, should we consider only the hypothesis H, or something more? It is known that some statisticians are of the opinion that good tests can be devised by taking into consideration only the hypothesis tested. But my opinion is that this is impossible and that, if satisfactory tests are actually devised without explicit consideration of anything beyond the hypothesis tested, it is because the respective authors *subconsciously* take into consideration certain relevant circumstances, namely, the alternative hypotheses that may be true if the hypothesis tested is wrong. However, it is rather difficult to discuss what an author may have in his mind subconsciously, or even consciously. The easier thing is to consider the situations which may present themselves when we are forced

[1] Since the publication of *Lectures and Conferences* in 1938, I have found that the first exact test of a statistical hypothesis was devised much earlier. In fact, this honor seems to belong to Laplace. In his paper, "Mémoire sur l'inclinaison moyenne des orbites des comètes," *Mémoires de l'Académie royale des Sciences de Paris,* Vol. VII, 1773 (see also *Oevres complètes* de Laplace, t. 8, Paris, 1891, pp. 279–321), Laplace deduced a test based on the exact distribution of the mean of a sample drawn from a "rectangular" distribution. Most readers of this book will be familiar with the fact that, when the sample size n is not too small, this distribution is very close to normal. Laplace gives the exact formula for the distribution and illustrates it on diagrams corresponding to several values of n. Curiously, while his formulae are correct, the diagrams are wrong and bear no resemblance to the normal law!

to select a test for a particular hypothesis H with nothing to base our device on except the hypothesis itself.

Suppose then that we have to test some hypothesis H, and that two different criteria T_1 and T_2 are suggested. Which of them should we use? What circumstances, referring to H and to nothing else, should influence our choice? I cannot think of all the suggestions that have been made, but I do remember seeing opinions that the criterion with the smaller standard deviation would be preferable.

Let us generalize this suggestion and consider more closely the tentative principle that the choice between possible criteria should be made on properties of their distributions as determined by H. This principle, call it Principle I, would obviously cover the question of the relative size of the standard deviations.

With regard to Principle I, I shall show that it is not sufficient for the choice. In fact, I shall prove that there may be two criteria having the following properties:

(i) Both have identical frequency distributions; and therefore, on the basis of Principle I alone, it will be impossible to choose between them.

(ii) Whenever one of these criteria has the most "improbable" values, thus "disproving" the hypothesis tested, the values of the other are just the most "probable" ones. This last circumstance will make it necessary to choose one of the criteria.

With the above situation in view, I shall mention another principle, to be called Principle II, which has been suggested by certain eminent workers in theoretical statistics: whenever you have two (or more) criteria, choose the one which, on the sample obtained, is less favorable to the hypothesis you test.

This principle implies, of course, that criteria could, and should, be chosen *after* the sample is drawn and analyzed.

I shall show that, if this principle is adopted, then it is useless to make any calculations with a view to testing hypotheses: given a certain amount of mathematical skill we shall be able to "disprove" any hypothesis on any sample.

The above two principles do not exhaust all the possibilities. There may be other principles that do not go beyond consideration of the hypothesis tested. For example, we may require of the functions T used as criteria some particular properties, e.g., that they should be symmetrical with respect to the random variables, etc. However, I cannot think of any such limitation that would seem reasonable. Therefore, without claiming that the two propositions which I am going to prove provide decisive evidence that it is absolutely impossible to make a rational choice of criteria without explicitly or tacitly considering hypotheses alternative to the one that is

being tested, I am inclined to think that this conclusion is highly probable. Anyhow, the two propositions do cover a certain range of possibilities and clear away certain popular misconceptions. They show, for instance, that an argument like "use T_1 rather than T_2 because its standard error is smaller" is not convincing. Let us now enter into details.

2. Insufficiency of Principle I. Consider a system of n random variables, $x_1, x_2, \cdots, x_n$, known to be independent and following the normal law

$$p_X(x_1 \cdots x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\Sigma(x_i-\mu)^2/2\sigma^2}, \tag{4}$$

where $\sigma > 0$ and μ are unknown constants. Suppose it is desired to test the hypothesis H that $\mu = 0$. This is known as Student's hypothesis. The generally accepted criterion to test H is the one invented by Student, namely, to calculate

$$z = \frac{\bar{x}}{s}, \tag{5}$$

where

$$\bar{x} = \frac{1}{n}\Sigma x_i, \qquad ns^2 = \Sigma(x_i - \bar{x})^2. \tag{6}$$

The probability law of z, if the hypothesis H be true, is given by

$$p_Z(z) = C(1 + z^2)^{-n/2}, \tag{7}$$

where

$$C^{-1} = \int_{-\infty}^{\infty} (1 + z^2)^{-n/2}\, dz = B[\tfrac{1}{2}(n-1), \tfrac{1}{2}]. \tag{8}$$

The hypothesis H is to be rejected whenever the value $|z'|$ of $|z|$ calculated for the sample is so large that

$$P\{|z| \geqq |z'|\} = 2\int_{|z'|}^{\infty} p(z)\, dz \tag{9}$$

is considered "small."

To prove the insufficiency of the Principle I as explained above I shall now define another criterion, depending on the quantity ζ, which will have the following properties:

1. If H be true, then the probability law of ζ is identical with that of z, so

$$p(\zeta) = C(1 + \zeta^2)^{-n/2}. \tag{10}$$

2. The absolute value of the product $|z\zeta|$ cannot exceed unity, i.e.,

$$|z\zeta| \leqq 1. \tag{11}$$

If the ζ criterion were used to test H, then this hypothesis would be rejected whenever $|\zeta|$ is large. In fact the large values of $|\zeta|$ are "improbable" whenever H is true. From (11) it follows that whenever $|\zeta|$ is large then $|z|$ must be small and conversely. Thus, whenever one of the alternative criteria z and ζ indicates that the hypothesis H should be rejected, the other is bound to protest that there is no reason for such rejection. This means that whenever one of the criteria has a large absolute value, we are compelled to choose the one whose verdict we shall respect. Principle I will not help us in the choice, because the probability laws of z and ζ are identical. This completes the proof of the insufficiency of Principle I.

In order to define ζ let us assume that the x_i are numbered in the order in which they are given by observation. Let

$$\bar{x}' = \frac{x_1 - x_2}{\sqrt{2n}} \tag{12}$$

and

$$ns'^2 = \sum_1^n x_i^2 - n\bar{x}'^2 = \tfrac{1}{2}(x_1 + x_2)^2 + \sum_3^n x_i^2. \tag{13}$$

The functions $\bar{x}'$ and s' thus defined will be called the quasi mean and the quasi standard deviation of the x_i. Now I shall prove *Proposition a*, namely that the ratio

$$\zeta = \frac{\bar{x}'}{s'} \tag{14}$$

has the properties 1 and 2 described above.

In order to prove 1, it is sufficient to show that the simultaneous probability law of $\bar{x}'$ and s' is identical with that of the ordinary mean $\bar{x}$ and standard deviation s.

If the hypothesis H be true, then $\mu = 0$ and

$$p_X(x_1, \cdots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\Sigma x_i^2/2\sigma^2}. \tag{15}$$

Let us introduce a new system of random variables, $y_1, y_2, \cdots, y_n$, connected with the x_i by the following formulas,

$$\left.\begin{aligned} x_1 &= y_1\sqrt{\frac{n}{2}} + y_2\sqrt{\frac{1}{2}}, \\ x_2 &= -y_1\sqrt{\frac{n}{2}} + y_2\sqrt{\frac{1}{2}}, \\ x_i &= y_i \qquad \text{for } i = 3, 4, \cdots, n. \end{aligned}\right\} \tag{16}$$

It will be noticed that

$$y_1 = \frac{x_1 - x_2}{\sqrt{2n}} = \bar{x}' \tag{17}$$

and is therefore identical with the quasi mean defined in equation (12). We shall return to this notation after a while. Furthermore,

$$y_2 = \frac{x_1 + x_2}{\sqrt{2}} \tag{18}$$

and having regard to (13) we shall have

$$s'^2 = \frac{1}{n}(y_2^2 + y_3^2 + \cdots + y_n^2). \tag{19}$$

The probability law of the y_i will be deduced from equation (15) following the steps indicated in my first lecture, namely,

$$p_Y(y_1, y_2, \cdots, y_n) = p_X(x_1, x_2, \cdots, x_n)\,|\,\Delta\,| \quad \text{(Eq. 25, page 21)} \tag{20}$$

where $|\,\Delta\,|$ is the Jacobian defined by equation (24) of page 21, and the x_i on the right-hand side should be expressed in terms of the y_i. Easy calculations give

$$p_Y(y_1, y_2, \cdots, y_n) \equiv p(\bar{x}', y_2, \cdots, y_n) = \frac{n}{(\sigma\sqrt{2\pi})^n} e^{-n(\bar{x}'^2 + s'^2)/2\sigma^2}, \tag{21}$$

where s'^2 stands for the sum of squares (19). Our next step consists in introducing still another system of variables, $u_1, u_2, \cdots, u_n$, one of which will be identical with $\bar{x}'$ and another with s'. We put

$$\left.\begin{aligned}
y_1 &= \bar{x}' = u_1, \\
y_2 &= \sqrt{n}u_2 \cos u_n \cos u_{n-1} \cdots \cos u_4 \cos u_3, \\
y_3 &= \sqrt{n}u_2 \cos u_n \cos u_{n-1} \cdots \cos u_4 \sin u_3, \\
y_4 &= \sqrt{n}u_2 \cos u_n \cos u_{n-1} \cdots \sin u_4, \\
&\cdot \\
&\cdot \\
&\cdot \\
y_n &= \sqrt{n}u_2 \sin u_n.
\end{aligned}\right\} \tag{22}$$

The range of variation of the new variables is determined by the following inequalities

$$\left.\begin{aligned} -\infty < u_1 < +\infty, \\ 0 < u_2, \\ 0 \leqq u_3 < 2\pi, \\ -\tfrac{1}{2}\pi < u_i < +\tfrac{1}{2}\pi, \quad i = 4, 5, \cdots, n, \end{aligned}\right\} \tag{23}$$

wherefore outside these limits the probability law of the u_i is identically equal to zero.

It will be easily seen that

$$u_2^2 = \frac{1}{n}(y_2^2 + y_3^2 + \cdots + y_n^2) \tag{24}$$

and later on we shall drop the notation u_1 and u_2, substituting for them $\bar{x}'$ and s' respectively. Easy calculations give for the Jacobian

$$\left|\frac{\partial(\bar{x}', y_2, \cdots, y_n)}{\partial(u_1, u_2, \cdots, u_n)}\right| = (\sqrt{n})^{n-1} u_2^{n-2} \cos u_4 \cos^2 u_5 \cos^3 u_6 \cdots \cos^{n-3} u_n \tag{25}$$

and it follows that

$$p_U(u_1, u_2 \cdots, u_n) = \left(\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\right)^n u_2^{n-2} e^{-n(u_1^2+u_2^2)/2\sigma^2} \cos u_4 \cos^2 u_3 \cdots \cos^{n-3} u_n. \tag{26}$$

In order to obtain the simultaneous probability law of u_1 and u_2 or, what comes to the same thing, of $\bar{x}'$ and s', we must integrate (26) for $u_3, u_4, \cdots, u_n$ from $-\infty$ to $+\infty$. Since the integrand differs from zero only within the limits shown in (23), and since these limits for $u_3, u_4, \cdots, u_n$ do not depend on the values of u_1 and u_2, we have at once that

$$p(u_1 u_2) = C_1 \left(\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\right)^n u_2^{n-2} e^{-n(u_1^2+u_2^2)/2\sigma^2}, \tag{27}$$

wherein

$$C_1 = \int \cdots \int_w \cos u_4 \cos^2 u_5 \cdots \cos^{n-3} u_n \, du_3 \, du_4 \, du_5 \cdots du_n \tag{28}$$

and the region of integration, w, is determined by

$$\left.\begin{aligned} 0 \leqq u_3 < 2\pi, \\ -\tfrac{1}{2}\pi < u_i < +\tfrac{1}{2}\pi \qquad \text{for } i = 4, 5, \cdots, n. \end{aligned}\right\} \tag{29}$$

Remembering that u_1 and u_2 are identical with $\bar{x}'$ and s' respectively, we have then

$$p(\bar{x}', s') = C_1 s'^{n-2} e^{-n(\bar{x}'^2+s'^2)/2\sigma^2}. \tag{30}$$

We see that the quasi mean and the quasi standard deviation as defined by (12) and (13) do follow a probability law identical with that of the ordinary mean $\bar{x}$ and standard deviation s of the x_i. In order to obtain the probability law of the ratio ζ we must now perform on equation (30) exactly the same operations that lead to the probability law of Student's z; and it is obvious that the probability law of ζ will be found to be identical with that of z. This proves the first part of the proposition.

Let us now prove part 2, namely, that $|z\zeta| \leqq 1$. For this purpose notice that, whatever be the real numbers a and b, we shall have

$$(a \pm b)^2 = a^2 \pm 2ab + b^2 \geqq 0 \tag{31}$$

and therefore

$$2|ab| \leqq a^2 + b^2. \tag{32}$$

It follows that for any real numbers a and b,

$$(a \pm b)^2 \leqq 2(a^2 + b^2). \tag{33}$$

If s is the ordinary standard deviation of the x_i and $\bar{x}$ their mean, then

$$ns^2 = \Sigma(x_i - \bar{x})^2 \geqq (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2. \tag{34}$$

On the other hand the definition of the quasi mean gives us

$$2n\bar{x}'^2 = (x_1 - x_2)^2 = [(x_1 - \bar{x}) - (x_2 - \bar{x})]^2 \tag{35}$$

and, from (33), we see that

$$2n\bar{x}'^2 \leqq 2[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2]. \tag{36}$$

Comparing (34) and (36) we find that

$$\bar{x}'^2 \leqq s^2, \tag{37}$$

an inequality between the squares of the quasi mean and of the ordinary standard deviation. From the definition of the quasi standard deviation (13) it follows that

$$\Sigma x_i^2 = n(s'^2 + \bar{x}'^2) = n(s^2 + \bar{x}^2). \tag{38}$$

Therefore

$$s'^2 + \bar{x}'^2 = s^2 + \bar{x}^2 \tag{39}$$

and, owing to (37),

$$\bar{x}^2 \leqq s'^2. \tag{40}$$

Multiplying (37) and (40) and dividing the resulting inequality by the product $s^2 s'^2$, we get

$$\left(\frac{\bar{x}^2}{s^2}\right)\left(\frac{\bar{x}'^2}{s'^2}\right) \leqq 1 \tag{41}$$

which is equivalent to $|z\zeta| \leqq 1$, or equation (11) of page 46. This fulfills

the proof of part 2 of *Proposition a*. Thus we have shown that Principle I by itself is not sufficient for a choice between alternative criteria that may be suggested for testing a given hypothesis.

3. Consequences of supplementing Principle I by Principle II. We shall now show that Principle I could not usefully be supplemented by Principle II. The combination of the two principles would read as follows: if there are several criteria for testing a given hypothesis H, all following the same probability law as determined by H, then the choice among them should be made *after* the sample is drawn and examined, and we should choose the test that appears to be the least favorable to H. We have already seen that if Student's hypothesis (page 46) be true, then Student's z is not the only function of the x_i following the familiar probability law (7). We shall now show that, whatever be the sample E' observed in a particular case, not all the x_i being equal to zero, it is possible to find a criterion, say ζ^0, which for this particular sample possesses the value $+\infty$ and which, in repeated sampling, follows exactly the same law as z and ζ discussed above. If we adopt both Principle I and Principle II, then we shall have to test Student's hypothesis using ζ^0; and this will lead to the rejection of the hypothesis. Thus in all cases, with the sole exception that all observed x_i are equal to zero, Student's hypothesis will have to be rejected, which shows that the combination of the two principles I and II is not a reasonable solution of the difficulty.

I shall now call the attention of the reader to the distinction between x_i' and x_i used below. The symbol x_i will mean, as before, the random variable following the law (15). On the other hand x_i' will denote a value of x_i observed in some particular case.

Proposition b.—Whatever be the sample

$$E' \equiv x_1', x_2', \cdots, x_n' \tag{42}$$

observed in a particular case, one at least of the x_i' being different from zero, it is possible to define a criterion ζ^0 which is represented by a function of the x_i and which has the following properties:

(i) The probability law of ζ^0, as determined by H, is the same as that of Student's z and that of ζ, equation (7), page 46.

(ii) The value $\zeta^0(E')$ of ζ^0, calculated for the sample E', is infinite.

It will be noticed that ζ^0 will have to be *adjusted to the sample* E' already observed. Therefore the values (42) will have to enter into the expression of ζ^0. They are constant numbers and will play the role of coefficients. On the other hand, ζ^0 will depend also on the random variables x_i.

Proof of part (i) *of Proposition b.*—Since the order in which the x_i are numbered is of no consequence, we may assume that $x_1', x_2', \cdots, x_m'$ are different

from zero, $m \leqq n$. Before defining ζ^0 we shall need the numbers $\alpha_1, \alpha_2, \cdots, \alpha_n$, which are connected with the x_i' by the n equations

$$\alpha_i = \frac{x_i'}{\sqrt{x_1'^2 + x_2'^2 + \cdots + x_n'^2}}, \qquad i = 1, 2, \cdots, n. \tag{43}$$

Obviously $\alpha_i \neq 0$ for $i = 1, 2, \cdots, m$, but $\alpha_i = 0$ for $i = m + 1, \cdots, n$; also

$$\sum_{i=1}^{n} \alpha_i^2 = 1. \tag{44}$$

Further steps consist first in defining a "pseudo mean" $\bar{x}''$ and a "pseudo S.D." s'' and then in making the identification

$$\zeta^0 \equiv \frac{\bar{x}''}{s''}. \tag{45}$$

Here the pseudo mean and pseudo S.D. are defined by

$$\bar{x}'' = \frac{\alpha_1 x_1 + \cdots + \alpha_n x_n}{\sqrt{n}} \tag{46}$$

and

$$s''^2 = \frac{1}{n} \Sigma x_i^2 - \bar{x}''^2. \tag{47}$$

It will be noticed that if $\alpha_i = 1/\sqrt{n}$ for $i = 1, 2, \cdots, n$, then the pseudo mean and pseudo S.D. become identical with the ordinary ones, $\bar{x}$ and s.

It will be sufficient to show the existence of a system of variables

$$v_1, v_2, \cdots, v_n,$$

whose elementary probability law as determined by H is

$$p_V(v_1, v_2, \cdots, v_n) = \frac{n}{(\sigma\sqrt{2\pi})^n} e^{-n(v_1^2 + s''^2)/2\sigma^2}, \tag{48}$$

wherein

$$v_1 = \bar{x}'' \quad \text{and} \quad ns''^2 = (v_2^2 + \cdots + v_n^2). \tag{49}$$

To show that $v_1, \cdots, v_n$ exist and that they possess the probability law (48) we introduce

$$\beta_k = \alpha_k[(\alpha_1^2 + \cdots + \alpha^2_{k-1})(\alpha_1^2 + \cdots + \alpha_k^2)]^{-\frac{1}{2}}$$

$$\text{for } k = 2, 3, \cdots, m \text{ and } \beta_k = 0 \text{ for } k > m. \tag{50}$$

Now, we relate $v_1, v_2, \cdots, v_n$ to $x_1, x_2 \cdots, x_n$ by the following system of equations:

$$\left.\begin{aligned}
x_1 &= \sqrt{n}\alpha_1 v_1 + \alpha_1\beta_2 v_2 + \alpha_1(\beta_3 v_3 + \beta_4 v_4 + \cdots + \beta_m v_m), \\
x_2 &= \sqrt{n}\alpha_2 v_1 - \frac{\alpha_1^2}{\alpha_2}\beta_2 v_2 + \alpha_2\beta_3 v_3 + \alpha_2(\beta_4 v_4 + \beta_5 v_5 + \cdots + \beta_m v_m), \\
x_3 &= \sqrt{n}\alpha_3 v_1 - \frac{\alpha_1^2 + \alpha_2^2}{\alpha_3}\beta_3 v_3 + \alpha_3\beta_4 v_4 + \alpha_3(\beta_5 v_5 + \cdots + \beta_m v_m), \\
&\vdots \\
x_k &= \sqrt{n}\alpha_k v_1 - \frac{\alpha_1^2 + \alpha_2^2 + \cdots + \alpha^2_{k-1}}{\alpha_k}\beta_k v_k + \alpha_k\beta_{k+1} v_{k+1} \\
&\quad + \alpha_k(\beta_{k+2} v_{k+2} + \cdots + \beta_m v_m),
\end{aligned}\right\} \tag{51}$$

for $k = 2, 3, \cdots, m$. In interpreting these equations, it is important to remark that, owing to the definition of m and β_k, if $m < n$, then

$$\beta_{m+1} = \beta_{m+2} = \cdots = \beta_n = 0. \tag{52}$$

If $m = n$, then equations (51) define the transformation completely. Otherwise, if $m < n$, we put

$$x_i = v_i \qquad \text{for } i = m + 1, \cdots, n. \tag{53}$$

With some algebraic reduction and the fact that $\alpha_1^2 + \cdots + \alpha_n^2 = 1$ (equation 44), it will be found that

$$v_1 = \frac{1}{\sqrt{n}}(\alpha_1 x_1 + \cdots + \alpha_n x_n) \equiv \bar{x}'' \tag{54}$$

and that

$$\begin{aligned}
(x_1^2 + \cdots + x_n^2) &= n v_1^2 + (v_2^2 + \cdots + v_n^2) \\
&= n v_1^2 + n s''^2.
\end{aligned} \tag{55}$$

The Jacobian $|\Delta| \equiv \left|\frac{\partial(x_1, x_2, \cdots, x_n)}{\partial(v_1, v_2, \cdots, v_n)}\right| = \sqrt{n}$, as is not difficult to work out from equations (51), (52), (53). From equation (55), and the value of the Jacobian, it follows by applying equation (25) of page 21 that if equation (15) is the simultaneous elementary probability law of $x_1, x_2, \cdots, x_n$, then that of $v_1, v_2, \cdots, v_n$ must be as written in equation (48).

Since equation (48) is of the same form as equation (21), and since formula (45) is similar to (14), it is clear that the steps required to deduce $p(\zeta^0)$ from (48) would be identical with those already shown in the deduction of $p(\zeta)$

from (21). This completes the proof that the criterion of ζ^0 has the property (i).

Proof of part (ii) *of Proposition b.*—We must now prove the other statement (ii) on page 51 concerning ζ^0; namely, we must prove that if in the expression for ζ^0 we substitute, instead of the random variables x_i, the particular observed values x_i' of (42) in terms of which the function ζ^0 has been defined, then the value $\zeta^0(E')$ of ζ^0 will be found infinite. Replacing x_i by x_i' in equation (46), and remembering that the coefficients α_i therein have already been defined by equation (43) in terms of the x_i', we easily find that the value of the pseudo mean calculated for the sample E' is

$$x''(E') = \frac{x_1'^2 + x_2'^2 + \cdots + x_n'^2}{n} > 0 \tag{56}$$

because at least one of the numbers x_i' is different from zero. Further, substituting x_i' for x_i in equation (47) to calculate the pseudo S.D. $s''(E')$, we find it to be zero. It follows from equation (45) that

$$\zeta^0(E') = \frac{x''(E')}{s''(E')} = \infty \tag{57}$$

and this completes the proof of part (ii) of *Proposition b.*

For the one particular sample E' already drawn, ζ^0 has the value ∞, but in repeated sampling it follows the same law as z and ζ.

It may be useful here to make the following remark. No number of examples is able to provide a proof of a general statement. On the other hand, the failure of a single example is sufficient to *disprove* any general statement. Our purpose here was to show that the principles I and II could not *generally* be applied for making a choice among criteria for testing hypotheses, and the validity of the proof does not suffer from the fact that we have limited ourselves to the consideration of one particular example.

As a matter of fact, it is easily seen how the above reasoning could be generalized, but such generalization would not produce any new relevant result.

4. General basis of the theory of testing statistical hypotheses. I shall finish this lecture by indicating what appears to be the general basis of the theory of testing statistical hypotheses. We must start by considering the situation in its most general form.

(i) When we desire to test a particular statistical hypothesis H_0, we imply that it may be wrong. E.g., if we try to test Student's hypothesis that $\mu = 0$, we admit the possibility that it may be wrong and that, therefore, μ may have some value other than zero. It will be seen that when-

ever we attempt to test a hypothesis we do admit, although perhaps subconsciously, that there are hypotheses that are contradictory or, in our terminology, alternative, to the one tested. There is no reason why these alternative hypotheses should not be considered explicitly when choosing an appropriate test.

(ii) Whenever we attempt to test a hypothesis we naturally try to avoid errors in judging it. This seems to indicate the right way of proceeding: when choosing a test we should try to minimize the frequency of errors that may be committed in applying this test.

Having in mind the above two points (i) and (ii) we may proceed further and discuss the kinds of errors we may commit in testing any given hypothesis H_0. It is easy to see that there are two kinds:

After having applied a test we may decide to reject the hypothesis H_0, when in fact, though we do not know it, it is actually true. This is called an *error of the first kind.*

After having applied a test we may decide not to reject the hypothesis H_0 (this may be described in short by saying that we "accept H_0") when in fact H_0 is wrong, and therefore some alternative hypothesis H' is true. This is called *an error of the second kind.*

The test adopted should control both kinds of errors. Now let us see what essentially is the machinery of any test, whatever be the principle on which it was chosen.

A test is nothing but a rule by which we sometimes reject the hypothesis tested and sometimes accept it (in the sense explained above), according to whether or not the observations available possess some properties specified by the rule. The observations are some n numbers, $x_1, x_2, \cdots, x_n$ the system of which could be represented by a point E in the n-dimensioned space W, having the x_i for the n coordinates. The point E and the space W are called the sample point and the sample space. Any rule specifying cases where we should reject the hypothesis tested is equivalent to a specification of the positions of E within W which, if arrived at by observation, lead to a rejection of H. These positions usually fill up a certain region, w, which is called the *critical region* or the *region of rejection.*

In conclusion we see that to choose a test for a statistical hypothesis H_0 we must choose a critical region w in the sample space W and make a rule of rejecting H_0 whenever E, as determined by observation, falls within w.

Let us illustrate this by an example. Consider the case where a sampled population is divided into n categories and we test the hypothesis that the probability of an individual falling within the ith category has some specified value p_i for $i = 1, 2, \cdots, n$. Denote by M the total number of

observations and by m_i the number of observations belonging to the ith category.

The generally accepted test of this hypothesis consists in rejecting it whenever

$$\chi^2 = \Sigma \frac{(m_i - Mp_i)^2}{Mp_i} \tag{58}$$

is "too large." What "too large" means is a subjective question, but there must be a more or less definite limit between values of chi-square that are "too large" and others that are not. Let χ_ε^2 denote this limit; and consider a space of $n - 1$ dimensions, the coordinates of any point being $m_1, m_2, \cdots, m_{n-1}$. As none of the m_i can be negative and their sum cannot exceed M, the sample space W will be composed of points E with all coordinates $m_1, m_2, \cdots, m_{n-1}$ being non-negative integers and satisfying the inequality

$$m_1 + m_2 + \cdots + m_{n-1} \leqq M. \tag{59}$$

It is easily seen that the rule of rejecting H_0 whenever $\chi^2 > \chi_\varepsilon^2$ is equivalent to considering the region w lying within W and outside the ellipsoid

$$\frac{(m_i - Mp_i)^2}{Mp_i} = \chi_\varepsilon^2 \tag{60}$$

as the critical region.

It is equally easy to see that any other test has a similar feature. For example, Student's test is equivalent to a rule of rejecting Student's hypothesis whenever the sample point falls within a circular hypercone with the axis

$$x_1 = x_2 = \cdots = x_n. \tag{61}$$

Having disposed of this we may go on to discuss the probabilities of errors. First of all: is it legitimate to discuss the probabilities of errors in testing statistical hypotheses? Isn't this equivalent to discussing the probabilities of hypotheses themselves, which would be useless? E.g., it would be useless to discuss the probability of Student's hypothesis because this would be the same as the probability of $\mu = 0$. As μ is an unknown constant, the probability of μ being equal to zero must be either $P\{\mu = 0\} = 0$ or $P\{\mu = 0\} = 1$ and, without obtaining precise information as to whether μ *is* equal to zero or not, it would be impossible to decide what is the value of $P\{\mu = 0\}$.

To this criticism the answer is the following. Undoubtedly, μ *is* an unknown constant and, as far as we deal with the theory of probability as described in my first two lectures, it is useless to consider $P\{\mu = 0\}$. On the other hand our verdict concerning the hypothesis tested, H_0, depends on the position of the sample point E, that is to say, on its coordinates, and these, according to our assumptions, are random variables. It follows that

our verdict is random and that there is no inconsistency in considering the probability of the verdict having this or that property, for example, of its being erroneous.

Consider the sample point E and any region w in the sample space. The probability of E falling within w may depend on the hypothesis that happens to be true. For example, if formula (4) represents the probability law of the x_i, and $\mu = 0$, then the probability of E falling within some particular region w may be 1/2. On the other hand if $\mu = 10$, say, the same probability may be equal to 0.0001. Therefore we shall agree to denote by $P\{E \in w \mid H\}$ the probability of E falling within w calculated on the assumption that the hypothesis H is true.

Now consider a hypothesis H_0 which we desire to test, and any region w which we have chosen to serve as critical region. What are the circumstances in which we commit an error of the first kind? They are: (i) the hypothesis tested is true; and (ii) the sample point E falls within the critical region w, whereupon H_0 is unjustly rejected. It follows that the probability of an error of the first kind must be calculated on the assumption that H_0 is true and, in fact, it is the probability

$$P\{E \in w \mid H_0\} \tag{62}$$

of E falling within w.

Now let us turn to errors of the second kind. For an error of the second kind to be committed it is necessary (and sufficient) that the hypothesis tested H_0 be wrong and that the sample point fail to fall within the critical region selected. But if H_0 is wrong, then some other admissible hypothesis H' must be true. Therefore, the probability of an error of the second kind is

$$1 - P\{E \in w \mid H'\}. \tag{63}$$

Obviously, instead of considering the probability of committing an error of the second kind, we may consider the probability of avoiding it, which is denoted by $\beta(w \mid H')$, so that

$$\beta(w \mid H') = P\{E \in w \mid H'\}. \tag{64}$$

$\beta(w \mid H')$ considered as a function of H' is described as the *power* (the power of detecting the falsehood of the hypothesis tested) of the region w with respect to the alternative hypothesis H'.

Any rational choice of a test must be made with regard to the properties of the power (64). Indeed, the values of the power $\beta(w \mid H)$ for a fixed region w and for a changing hypothesis H (which in particular may be H_0, the one we desire to test) give no more and no less than a complete description of the properties of the test based on the critical region w. In fact, what could be called "the properties of a test?" To know the proper-

ties of a test can mean nothing but to know (i) how frequently this test will reject the hypothesis H_0 tested, when it is true; and (ii) how frequently it will disprove H_0 when H_0 is wrong. That is exactly what the values of the function $\beta(w \mid H)$ tell us. Without knowing the properties of $\beta(w \mid H)$, we cannot very well say that we know the properties of a test based on w. And just these properties of the power seem to be the proper rational basis for choosing a test.

For example, by considering the power of Student's test, it is possible to show that this test has the following properties, which put it above any other test that may be suggested.

1. The probability of rejecting the hypothesis H_0 that $\mu = 0$ is always greater when the hypothesis H_0 is wrong than in cases when H_0 is true. This property is described by the adjective "unbiased" attached to the test possessing the property.

2. Any other unbiased test, if it leads to the same frequency of errors of the first kind, will less frequently detect the falsehood of the hypothesis H_0 when H_0 is in fact wrong.

The responsibility for the above concepts and for the resulting theory of testing statistical hypotheses is borne jointly by Egon S. Pearson and the present writer. Our first paper [2] on the subject was published in 1928, over twenty years ago. However, it took another five years for the basic idea of a rational theory to become clear in our minds.[3] Thereafter, the work became easier and within a short time we were joined by a number of colleagues.[4]

Bare statements of principles are never clear unless the principles are illustrated in full detail with examples. It would be most satisfactory if the use of the concepts described above could be illustrated with examples which are both easy and of practical importance. Unfortunately, it is very difficult to satisfy both conditions at the same time. One must choose between the illustrativeness of an example which involves a certain artificiality and the practical importance of a test which involves technical difficulties in dealing with the problem. Faced with the necessity of choosing between the two alternatives, the writer felt that the readers of this book would be best served by a simple illustrative example, even though it is somewhat artificial.

We will imagine an early stage in the study of a pair of genes, the domi-

[2] J. Neyman and E. S. Pearson: "On the use and interpretation of certain test criteria for purposes of statistical inference." *Biometrika,* Vol. 20-A (1928), pp. 175–240 and 264–299.

[3] J. Neyman and E. S. Pearson: "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc.,* London, Vol. 231A (1933), pp. 289–337. Recently, a systematic elementary presentation of the theory was given in the author's *First Course on Probability and Statistics* already quoted.

[4] See: *Statistical Research Memoirs,* Vol. I (1936), Vol. II (1938).

nant gene to be called G, the recessive g. We imagine that it is more or less taken for granted that the mating of the organisms carrying these genes is non-assortative (i.e., that the genetical composition of one mate is independent of that of the other mate) and is of uniform fertility. Contrary to this general belief, a geneticist suspects that the recessive types gg do not participate in the reproduction. This suspicion is not based on any trials but on some analogies, and, in preparing for a meeting at which the genes G and g are to be discussed, the geneticist is somewhat hesitant whether or not to come out with his doubts. Before deciding, he wishes to take into account the results of two independent experiments performed for other purposes, but involving genes G and g. Both experiments had the same pattern. In each case two hybrids $Gg \times Gg$ were crossed, giving a generation of progeny which we shall denote by F_1. Next the F_1 individuals were allowed to mate without interference, producing the second generation F_2. Finally the F_2 individuals were allowed to mate without interference and they produced the third generation F_3. Since the two experiments were carried out for purposes not connected with genes G and g, the records of the experiments appear to be fragmentary as far as the genes G and g are concerned. In fact, the only information concerning these genes in the first experiment is that the F_2 generation was composed of $n_1 = 8$ individuals and that among them there were exactly x_1 recessives gg. Further, the records of the second experiment show only that the F_3 generation was composed of $n_2 = 10$ individuals and that among them there were exactly x_2 recessives gg. The values of the four numbers n_1, x_1 and n_2, x_2 must now be used by the geneticist to make up his mind whether or not to voice doubts about the non-assortative character of mating. Every human action is subject to error, and therefore the geneticist would not mind being in error from time to time. However, he is inclined to lay down rules for his behavior so as to control the frequency of errors. First, in cases where some established hypotheses are true, he would like to voice doubts of these hypotheses only rarely, say with a frequency not exceeding a selected number α, perhaps $\alpha = .1$ or $\alpha = .05$ or the like. Another requirement which the geneticist lays down for his behavior is that, in cases where some hypothesis H_2, alternative to the established hypothesis H_1, is true, then he wants his rule to lead him to protest as frequently as is humanly possible.

Applying these two principles to the case of the genes G and g, the geneticist notices that n_1 and n_2 are sure numbers while x_1 and x_2 are random variables whose particular values are determined by the two experiments. Let H_1 and H_2 denote the two hypotheses under consideration. Namely, H_1 asserts that, with respect to genes G and g, the mating is non-assortative with uniform fertility and H_2 asserts that the non-assortativeness and uniform fertility apply only to dominant and hybrid types GG

and Gg, but that the recessives gg do not participate in the reproduction. We will assume for simplicity that the geneticist admits the possibility of only these two hypotheses H_1 and H_2.

On either hypothesis, the random variables x_1 and x_2 are capable of assuming all the 99 different combinations of integer values $x_1 = k_1$ and $x_2 = k_2$, with $k_1 = 0, 1, 2, \cdots, 8$ and $k_2 = 0, 1, 2, \cdots, 10$. Thus the sample space W is composed of 99 points with coordinates (k_1, k_2). Easy calculations give the probability that the sample point $E = (x_1, x_2)$ will assume the position (k_1, k_2). Namely, on the hypothesis H_1 we have, say,

$$p(k_1, k_2 \mid H_1) = P\{(x_1 = k_1)(x_2 = k_2)\} = C_{n_1}^{k_1} C_{n_2}^{k_2} (\tfrac{1}{4})^{k_1+k_2} (\tfrac{3}{4})^{n_1+n_2-k_1-k_2}. \tag{65}$$

On the hypothesis H_2 we have

$$p(k_1, k_2 \mid H_2) = P\{(x_1 = k_1)\,(x_2 = k_2) \mid H_2\} = C_{n_1}^{k_1} C_{n_2}^{k_2} (\tfrac{1}{9})^{k_1} (\tfrac{8}{9})^{n_1-k_1} (\tfrac{1}{16})^{k_2} (\tfrac{15}{16})^{n_2-k_2}. \tag{66}$$

Tables I and II give the numerical values of these probabilities for all combinations of $k_1 = 0, 1, 2, \cdots, 8$ and $k_2 = 0, 1, 2, \cdots, 10$ in so far as these probabilities are not too small. Upon adding all the entries in Table I the reader will obtain the total .998. Thus the probability is approximately .002 that the sample point E will occupy any position in

TABLE I

Joint probability distribution of x_1 and x_2, $P\{(x_1 = k_1)(x_2 = k_2) \mid H_1\}$, as determined by the hypothesis H_1

k_2	k_1							
	0	1	2	3	4	5	6	7
0	.006	.015	.018	.012	.005	.001	.000	.000
1	.019	.050	.058	.039	.016	.004	.001	.000
2	.028	.075	.088	.058	.024	.006	.001	.000
3	.025	.067	.078	.052	.022	.006	.001	.000
4	.015	.039	.046	.030	.013	.003	.001	.000
5	.006	.016	.018	.012	.005	.001	.000	.000
6	.002	.004	.005	.003	.001	.000	.000	.000
7	.000	.001	.001	.001	.000	.000	.000	.000
8	.000	.000	.000	.000	.000	.000	.000	.000

TABLE II

Joint probability distribution of x_1 and x_2, $P\{(x_1 = k_1)(x_2 = k_2) \mid H_2\}$, as determined by H_2

k_2	k_1					
	0	1	2	3	4	5
0	.204	.204	.089	.022	.003	.000
1	.136	.136	.060	.015	.002	.000
2	.041	.041	.018	.004	.001	.000
3	.007	.007	.003	.001	.000	.000
4	.001	.001	.000	.000	.000	.000
5	.000	.000	.000	.000	.000	.000

the sample space for which the entry in Table I is zero or is not listed at all. The same probability for Table II is equal to .004.

Consider now the problem of selecting the combinations of values of x_1 and x_2 such that, if any one of these combinations is determined by the two experiments, then the geneticist would consider it advisable to reject the hypothesis H_1. In the terminology of this lecture, the problem is that of selecting the critical region w_0 for testing the hypothesis H_1 against the set Ω of admissible hypotheses which, in this case, includes H_1 and H_2 only. The principles which the geneticist laid down for his choice are exactly those determining the best critical region for testing H_1 against Ω. The first of these principles is that the region w_0 be one of those regions w for which

$$P\{E \in w \mid H_1\} \leqq \alpha. \tag{67}$$

The second principle is that, if w_0 is the selected region and w any other region such that

$$P\{E \in w \mid H_1\} \leqq P\{E \in w_0 \mid H_1\}$$

then

$$P\{E \in w \mid H_2\} \leqq P\{E \in w_0 \mid H_2\}.$$

The construction of the critical region w_0 having this property is easily accomplished by the following simple rule, the validity of which will be proved in general, for any number of discrete observable random variables $X_1, X_2, \cdots, X_n$.

Denote generally by $e_1, e_2, \cdots, e_n, \cdots$ all possible positions of the sample point E as may be determined by some observations. Let further $p(e_k \mid H_1)$ and $p(e_k \mid H_2)$ denote the probabilities determined by the

hypotheses H_1 and H_2, respectively, that E will coincide with e_k. Here some of the probabilities $p(e_k \mid H_i)$, $i = 1, 2$, may be zero while others are positive. For each point e_k for which $p(e_k \mid H_2) > 0$ define the ratio

$$R(e_k) = \frac{P(e_k \mid H_1)}{P(e_k \mid H_2)}.$$

Lemma. IF a IS A POSITIVE NUMBER AND w_0 A REGION IN THE SAMPLE SPACE SUCH THAT IT INCLUDES ALL POINTS e_k FOR WHICH $R(e_k) < a$ AND NONE OF THOSE POINTS e_m FOR WHICH $R(e_m) > a$, THEN, WHATEVER BE ANY OTHER REGION w SUCH THAT

$$P\{E \in w \mid H_1\} \leqq P\{E \in w_0 \mid H_1\}, \tag{68}$$

NECESSARILY

$$P\{E \in w \mid H_2\} \leqq P\{E \in w_0 \mid H_2\}.$$

If the regions w_0 and w are contemplated as critical regions for testing H_1, then $P\{E \in w \mid H_i\}$ is the probability that H_1 will be rejected using w in those cases when the true hypothesis is H_i. Thus $P\{E \in w \mid H_1\}$ is the probability of an erroneous rejection of H_1 (that is, rejection when H_1 is true, or the probability of an error of the first kind). On the other hand, $P\{E \in w \mid H_2\}$ is the probability of rejecting H_1 when the true hypothesis is H_2, i.e., it is the power of the test based on w. This property of w_0 may be described verbally by stating that out of all critical regions w which control the errors of the first kind as well as w_0 or better, the critical region w_0 has the greatest power.

In proving the Lemma it will be convenient to use the following notation. Let u be some region in the sample space and let

$$e_{k_1}, e_{k_2}, \cdots, e_{k_m}$$

be all the possible positions of the sample point E which fall within the region u. Then the probability $P\{E \in u \mid H_i\}$ that the sample point will fall within u is given by the sum

$$P\{E \in u \mid H_i\} = \sum_{j=1}^{m} p(e_{k_j} \mid H_i).$$

It will be convenient to denote this last sum simply by $\sum_u p(e \mid H_i)$.

With this notation, the inequality (68) can be rewritten as

$$\sum_{w_0} P(e \mid H_1) \geqq \sum_{w} P(e \mid H_1)$$

and it follows that, say,

$$\Delta(H_1) = \sum_{w_0} p(e \mid H_1) - \sum_{w} p(e \mid H_1) \geqq 0. \tag{69}$$

The two regions w_0 and w may have a common part which we will denote by v. Should there be no common part of w_0 and w, then v will stand for the "empty" set of points. In any case we may write that

$$\left.\begin{aligned} w_0 &= (w_0 - v) + v, \\ w &= (w - v) + v, \end{aligned}\right\} \tag{70}$$

and it is clear that every point in $w - v$ lies *outside* of w_0.

Obviously $\Delta(H_1)$ can now be rewritten using the summation over the regions $w_0 - v$ and $w - v$,

$$\Delta(H_1) = \sum_{w_0 - v} p(e \mid H_1) - \sum_{w - v} p(e \mid H_1) \geqq 0. \tag{71}$$

The region $w_0 - v$ contains only points e_k which are interior to w_0. Because of the definition of w_0, for each of these points

$$p(e \mid H_1) = R(e)p(e \mid H_2) \leqq ap(e \mid H_2). \tag{72}$$

Therefore, say

$$\Delta' = a \sum_{w_0 - v} p(e \mid H_2) - \sum_{w - v} p(e \mid H_1) \geqq \Delta(H_1) \geqq 0. \tag{73}$$

Since each point e belonging to $w - v$ lies outside of w_0, the definition of w_0 implies that for each such point

$$p(e \mid H_1) = R(e)p(e \mid H_2) \geqq ap(e \mid H_2). \tag{74}$$

Therefore

$$\Delta(H_2) = a \sum_{w_0 - v} p(e \mid H_2) - a \sum_{w - v} p(e \mid H_2) \geqq \Delta' \geqq \Delta(H_1) \geqq 0. \tag{75}$$

Since a is a positive number, it follows that

$$\sum_{w_0 - v} p(e \mid H_2) \geqq \sum_{w - v} p(e \mid H_2). \tag{76}$$

Adding to both sides of this inequality the same sum $\sum_{v} p(e \mid H_2)$, we obtain the desired result, namely,

$$P\{E \in w_0 \mid H_2\} = \sum_{w_0} p(e \mid H_2) \geqq \sum_{w} p(e \mid H_2) = P\{E \in w \mid H_2\}. \tag{77}$$

This completes the proof of the Lemma.

It follows from the Lemma that the operations necessary for determining a best critical region for testing H_1 with respect to a single alternative hypothesis H_2 are the following.

(i) Compute the ratio $R(e)$ for all possible sample points.

(ii) Renumber the possible sample points in order of the magnitude of the corresponding ratios $R(e)$, beginning with the smallest $R(e_1)$, so that

$$R(e_1) \leqq R(e_2) \leqq \cdots \leqq R(e_{k-1}) \leqq R(e_k) \leqq \cdots. \tag{78}$$

(iii) Include e_1 in the critical region w_0 and also as many of the following points, e_2, e_3, $\cdots$, e_m as possible without impinging upon the condition that the probability determined by H_1 of the sample point E falling within w_0 does not exceed α,

$$P\{E \in w_0 \mid H_1\} = \sum_{i=1}^{m} p(e_i \mid H_1) \leqq \alpha. \tag{79}$$

Returning to the problem of testing the hypothesis H_1 concerned with non-assortative mating and uniform fertility, we could proceed in two slightly different ways. One of these consists in computing the ratios $R(e)$ numerically as indicated in step (i). The disadvantage of this method is that it is somewhat cumbersome and involves ratios of numbers which are so small that they are not recorded in Tables I and II.

The other method is to compute the formula for $R(e)$. We have, say

$$\left.\begin{aligned} R(e) = R(k_1, k_2) &= \frac{p(k_1,k_2 \mid H_1)}{p(k_1,k_2 \mid H_2)} \\ &= \frac{3^{n_1+n_2-k_1-k_2}9^{n_1}16^{n_2}}{4^{n_1+n_2}8^{n_1-k_1}15^{n_2-k_2}} \\ &= C(\tfrac{8}{3})^{k_1}5^{k_2} = CR'(k_1,k_2). \end{aligned}\right\} \tag{80}$$

where, for the sake of brevity, the letter C is used to denote the numerical factor

$$C = \frac{3^{n_1+n_2}9^{n_1}16^{n_2}}{4^{n_1+n_2}8^{n_1}15^{n_2}} \tag{81}$$

which is independent of k_1 and k_2. It is obvious that instead of ordering the points e in the order of magnitude of $R(e)$, we may order them in the order of magnitude of $R'(k_1k_2)$ or, since this is even more convenient, in the order of magnitude of, say

$$\left.\begin{aligned} r(k_1k_2) = \log_{10} R'(k_1,k_2) &= k_1 \log_{10} (\tfrac{8}{3}) + k_2 \log_{10} 5 \\ &= k_1(.42597) + k_2(.69897). \end{aligned}\right\} \tag{82}$$

Now it is obvious that the first point to be included in w_0 is the one corresponding to $k_1 = k_2 = 0$. The next most desirable point is $k_1 = 1, k_2 = 0$, etc. Table III gives the ordering of points (k_1, k_2) as indicated in step (ii), the corresponding values of $r(k_1, k_2)$, the corresponding probability determined by H_1 and H_2 and the cumulative sums of these probabilities. The most interesting columns in Table III are columns (5) and (7). Column (5) gives the probabilities determined by H_1 that the point E to

TABLE III

Steps (ii) and (iii) in determining w_0

(1) Order of the point e_i	(2) Coordinates k_1, k_2	(3) $r(k_1, k_2)$	(4) $p(e_i \mid H_1) = p(k_1, k_2 \mid H_1)$	(5) $\sum_{j=1}^{i} p(e_j \mid H_1)$	(6) $p(e_i \mid H_2) = p(k_1, k_2 \mid H_2)$	(7) $\sum_{j=1}^{i} p(e_j \mid H_2)$
e_1	0, 0	.00000	.006	.006	.204	.204
e_2	1, 0	.42597	.015	.021	.204	.408
e_3	0, 1	.69897	.019	.040	.136	.544
e_4	2, 0	.85194	.018	.058	.089	.633
e_5	1, 1	1.12494	.050	.108	.136	.769

be determined by observing x_1 and x_2 will fall within the critical region w_0 including only the point e_1, or the two points e_1 and e_2, or three points e_1, e_2, e_3, etc. These probabilities, then, are the probabilities of wrongly rejecting the hypothesis H_1 when it is in fact true, corresponding to an increasing critical region w_0. For example, if the geneticist decides that he should not raise false doubts concerning hypotheses more often than five times in a hundred when such hypotheses are true, then his critical region should include only three points (0, 0), (1, 0) and (0, 1) with the resulting probability of an error of the first kind equal to .040. Should this be his decision, then the probability of detecting that H_1 is false when the true hypothesis is H_2 (or the power of the test), is .544. It is found in column (7) of Table III.

However, the geneticist may compromise on the probability of the error of the first kind equal to .058, or even .108. Then his chances of detecting the falsehood of H_1 when the true hypothesis is H_2 will be .633 or .769, respectively.

Whichever critical region is finally adopted, including any number of the first points e_i ordered according to the value of $r(k_1, k_2)$, the Lemma guarantees that the power of the resulting test cannot be improved by using any other critical region which controls the errors of the first kind to the same (or better) level as the region chosen.

Suppose now, that the values of x_1 and x_2 that were actually observed are $k_1 = 2$, $k_2 = 0$. It follows from the foregoing that, if the geneticist does not insist on the probability of an error of the first kind being less than .058, he should go ahead and voice his doubts of the hypothesis H_1 of non-assortativeness of mating and of uniform fertility. In taking this

step he should be aware that the above analysis does not contribute anything about the falsehood or correctness of the particular genetical hypothesis H_1. In fact, no test can reveal any definite information about any statistical hypothesis if the values of the observable random variables which are possible under this hypothesis are also possible under some alternative one. All the geneticist can be certain about is that, if his attitudes towards statistical hypotheses are consistently governed by analyses such as the one described, with a fixed value of α, then, in the long run, the relative frequency of his raising doubts concerning hypotheses, when such doubts are unjustified, will not exceed α. Moreover, he can also be sure that, in cases when the hypothesis tested H_1 is wrong, the chance of the above method detecting the falsehood of H_1 is as good as or better than that corresponding to any other method insuring the same level of control of errors of the first kind.

The reader may be interested in considering critical regions for testing H_1 against H_2 other than the ones suggested in Table III. For example, the reader may wish to compute the probability of error of the first kind and the power of critical regions whose selection is based on the probability distribution of x_1 and x_2 determined by H_1. Upon examining Table III one might perhaps suggest the critical region w' composed of all possible sample points e for which

$$p(e \mid H_1) \leq .001 \tag{83}$$

or, perhaps the critical region w'' composed of all points such that

$$p(e \mid H_1) \leq .005, \tag{84}$$

etc. It will be seen that regions of this kind will control errors of the first kind to levels comparable to those of regions w_0, suggested in Table III. However, there will be a marked difference between the two kinds of tests in their power to detect the falsehood of H_1 when the true hypothesis is H_2.

CHAPTER II

Some Controversial Matters Relating to Agricultural Trials

Part 1. Randomized and Systematic Arrangements of Field Experiments

(The contents of this lecture are based on a conference at the Cosmos Club, Washington, D. C., held April 7, 1937, under the chairmanship of Dr. Frederick F. Stephan and also on some sections of papers published in the *Supplement to the Journal of the Royal Statistical Society*, Vol. 2, 1935.)

I am going to speak on a very controversial question: Can systematically arranged agricultural trials be treated with any success by means of mathematical statistics? Two eminent statisticians who are also experts in agricultural experimentation disagree drastically on the answer and each of them has a number of supporters. One of these scientists, Professor R. A. Fisher, claims that, in arranging field experiments systematically, we are bound to obtain all sorts of biases in our estimates and thus to ruin the statistical tests. The other scientist is "Student" who can be considered, and rightly so, the father of statistical work in agricultural experimentation. He does not deny that the formulas usually applied to estimate the experimental standard error in both randomized and systematic trials are in the latter case somewhat biased and tend to overestimate the error. But it is his claim that the actual accuracy of a systematic experiment is usually greater than that of a randomized one. In his opinion, too high an estimate of the standard error is not especially important, since it keeps the experimenter on the safe side.

Members of the present audience who are familiar with the material of my first two lectures are aware that the answer to the question must be both empirical and subjective. Since the application of formulas of mathematical statistics to the results of agricultural trials presumes the existence of a mathematical model of these experiments, the question under consideration reduces to one of whether or not the correspondence between the model and what happens in actual practice is sufficiently accurate. This question is exactly similar to the one mentioned in my second lecture (page 23): "Can the formulas of plane geometry be applied to measure this or that area on the surface of the earth?" Another similar problem (page 28) is whether or not formulas deduced from the Poisson law of

frequency can be successfully used to estimate the probability that a colony on a Petri plate is produced by a single individual.

The empirical character of the answer arises from the fact that the answer involves trials in conditions of actual practice. The subjective character is unavoidable, because, after we have the results of the trials and also the corresponding theoretical deductions from their mathematical model, we must judge whether the agreement is or is not satisfactory. One of the ways by which the insufficiency of plane geometry may be revealed consists in subdividing an area of the type it is desired to measure into several suitable partial ones and in measuring each of the parts. If the measure of the whole appears to be very different from the sum of the measures of its parts, then we would say that the assumption that the area measured is plane is too crude. But it will be up to us to decide whether the disagreement between the two measures *is* actually large or not, and in this respect personal opinions vary.

Having this in view, I am going to give a short account of the work recently done by Mr. C. Chandra Sekar in the Department of Statistics, University College, London. This provides the objective empirical part of the answer to the question discussed by Fisher and Student. The results that I shall describe are of the same character as those contained in my second lecture (pp. 30–41): on the one hand you will see figures representing frequencies of various results, as predicted from the mathematical models of the agricultural trials, and on the other, the frequencies actually observed. If the agreement between the two is judged satisfactory, the conclusion will be that there is no special harm in arranging the experiments systematically. If, on the other hand, you find that the agreement is bad, you will require an alteration either of the mathematical model or of the experimental design. For example, you may decide to randomize your trials.

Now I must enter into details and describe the experiments that I have in mind. I shall deal with experiments of a very common type in which the plots are rather narrow, long rectangles all arranged in one row. They are combined into a few blocks and within each block all the compared agricultural objects (varieties or treatments) are distributed in one way or another. This is the general description. If we add to this some details on the way the objects are distributed within the blocks, we shall obtain the full description of the two types of arrangements under discussion.

One of these is the so-called arrangement in *randomized blocks*. In this arrangement, as you know, each of the objects is repeated in each of the blocks the same number of times, e.g. once, and the order in which the objects occur within each block is determined by random sampling. If the number of compared objects is four and they are denoted by A, B, C, D,

then in a randomized block experiment we may find the following distribution of objects on the successive plots.

Block I	Block II	Block III	Block IV
A C D B	B C A D	C D A B	B A C D

This is one type of arrangement and we know the formula by which we can calculate the estimates of the true difference between the mean yields which any two of the objects compared, say A and B, are able to give if sown over the whole field. It is the difference between the means $x_A - x_B$ of the observed yields. Also, we know how to calculate an unbiased estimate s^2 of the variance of our result. Owing to the fact that the observations referring to one block are mutually dependent (e.g., if the object A got the best of the four plots, then the object B *must* have gotten one of the poorer plots), the further theory is not entirely clear.[1]

It is probable, however, that the application of the t test gives results very much in accordance with its theory: i.e., the hypothesis tested, namely, that there is no difference between the mean yields of the objects compared, is rejected both when it is true and when it is false with relative frequencies in good accord with the mathematical tables.

Many practical agriculturists find that the objects compared are not always satisfactorily distributed over the field if the distribution is left to chance. For example, they would object to the variety B being sown twice on adjoining plots. In their opinion, the conditions in which the particular objects are compared should be as equal as possible, and they think that this is best attained by some systematic distribution of the objects, such as the following.

Block I	Block II	Block III
A B C D	A B C D	A B C D

etc.

Frequently, though not always, a field experiment arranged in the above manner is treated statistically by means of the formulas mentioned above,

[1] J. Neyman with cooperation of K. Iwaszkiewicz and S. Kolodziejczyk: "Statistical problems in agricultural experimentation." *Supplement to the Roy. Stat. Soc.*, Vol. 2 (1935), pp. 107–180.

See also Michael D. McCarthy: "On the application of the z-test to randomized blocks." *Annals of Math. Stat.*, Vol. 10 (1939), pp. 337–359.

formulas meant for randomized block experiments. There is no doubt that from the point of view of theory this procedure is wrong. The theory of randomized blocks assumes specifically that the blocks *are* randomized and its validity is easily shown to depend on this assumption. However, it is a question, not of the fact that discrepancies do arise from the disregard of this condition, but of the size of these discrepancies between theory and practice.

The above systematic arrangement is very popular in Poland. I spent much time and wasted much paper trying to persuade practical experimenters to randomize their blocks, but with disappointing success. Then the thought occurred to me that the agreement between theory and practice may be attained not only by altering the practice, but also by adjusting the theory. Consequently, I produced a paper [2] giving a statistical theory of the agricultural trials arranged systematically.[3]

The general lines are as follows. It is assumed that the natural level of fertility along a field may be adequately represented by a parabola of some not very high order, say the fourth. If u denotes the coordinate of the center of any of the plots, starting from the left, so that

$$u = 1, 2, \cdots, N, \tag{1}$$

then the true yield of A, if it were tested on the uth plot would be

$$A(u) = A + bu + cu^2 + du^3 + eu^4, \tag{2}$$

where A is a term depending on the object A (treatment or variety), and b, c, d and e are unknown coefficients. The symbol A is used here to signify both the thing being tested (treatment or variety), and the true value (as the yield) of the thing being tested. Experience has shown, however, that confusion does not arise, and in fact the symbolism is a very convenient one. The true yield of the object B, if it were sown on the same plot would be given by

$$B(u) = B + bu + cu^2 + du^3 + eu^4, \tag{3}$$

where B depends on the object B but the other constants b, c, d, and e are the same as in equation (2). Similar relations are written for C, D, etc., b, c, d, and e being the same for all.

In actual experiments we do not obtain what we call the 'true" yields. What we obtain is the sum of the true yield plus an experimental error,

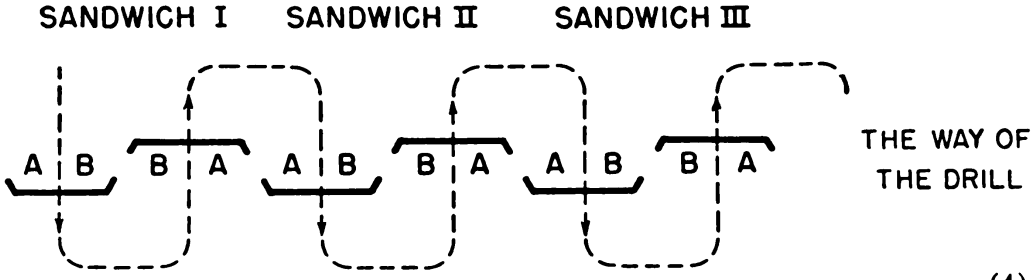
[2] J. Neyman: *The theoretical basis of different methods of testing cereals, Part II: The method of parabolic curves.* K. Buszczynski and Sons, Ltd., Warsaw, 1929, 48 pp.

[3] In more recent times my formulae were refound by A. Hald. See A. Hald: *The decomposition of a series of observations.* G. E. C. Gads Forlag, Copenhagen, 1948, 134 pp.

due to various factors, such as inaccuracies in measuring plots, in treatment, damage by birds, etc. My assumption was that these experimental errors on particular plots are independent of each other. I then applied the Markoff [4] theorem to get estimates of the differences, $B - A$, $C - A$, etc., and of their respective variances.

If the assumptions are granted, the theory is correct. It certainly corresponds more exactly to the practice of systematic experiments than the theory of randomized blocks does, but for a long time there was no answer to the question of what this correspondence meant in figures. Now some numerical evidence is available indicating that the theory does correspond to what happens in practice, at least in one particular type of systematic arrangement called half drill strip.

This experimental design was invented by Dr. E. S. Beaven [5] who used it with great success while breeding his renowned varieties of barley. The half-drill-strip experiments are designed to compare only two objects, say two varieties, A and B. The varieties are sown in long narrow plots, half the drill sowing A, the other half B. The varieties are repeated in a systematic order as follows.



(4)

Four consecutive plots form what is called a sandwich, two half drill strips with B, sown in opposite directions, are enclosed between two with A, also sown in opposite directions. These sandwiches obviously correspond to blocks, but the blocks are not randomized.

It will be useful to distinguish between two possible methods of randomizing the blocks of four plots to be occupied by two varieties only. One would be a totally unrestricted randomization, allowing arrangements like

AABB, ABAB, ABBA, BAAB, BABA, BBAA. (5)

The second kind of randomizing would consist in randomizing the sand-

[4] See F. N. David and J. Neyman: "Extension of the Markoff theorem on least squares." *Statistical Research Memoirs*, Vol. II (1938), pp. 105–116.

[5] E. S. Beaven: "Trials of new varieties of cereals." *Jr. of the Ministry of Agriculture*, Vol. 29 (1922), nos. 4 and 5, pp. 1–28, 436–444.

wich. This would admit only two arrangements of the block, either $ABBA$ or $BAAB$, and the choice between them should be based on some random experiment such as tossing a coin.

If the sandwiches are randomized as just described, and if x_i denotes the difference between the sum of the two yields of A and the two yields of B observed on the ith sandwich, then the ordinary theory of randomized blocks is applicable to the x_i. But this is not so certain with respect to a systematic arrangement like (4). Of course, the arrangement (4) may be treated by the method of parabolic curves described above. It is a matter of an easy adjustment of a few formulas and of preparing tables to facilitate the calculations. But here again we come to the question of whether or not the scheme underlying the method of parabolic curves corresponds with sufficient accuracy to what happens in practice.

I shall now discuss the question of the empirical data needed for deciding whether or not any particular mathematical model corresponds to the experiments.

When comparing any two objects A and B, of which A is some established standard, we may desire to obtain evidence that B is better than A. This reduces to the test of the statistical hypothesis H_0 that the true average yield $\bar{B}$ of B if sown on the whole field, does not exceed that of A, say $\bar{A}$. That is, H_0 is the hypothesis that

$$\bar{B} - \bar{A} \leqq 0. \tag{6}$$

Whichever one of the mathematical schemes described is applied, the test of H_0 consists (i) in calculating the estimate of $\Delta = \bar{B} - \bar{A}$, say $\bar{x}$, (ii) in calculating the estimate s^2/n of the variance of $\bar{x}$, and (iii) in referring the quotient $t = \bar{x}/(s/\sqrt{n})$ to Fisher's table of t. If the observed value of t exceeds the value tabled t_α, corresponding to some small value of P, say 0.05 or 0.01, then the hypothesis H_0 is rejected and we consider that we have "evidence" of B being able to give average yields greater than A.

The whole question under discussion, i.e., whether or not the field trials *must* be randomized, whether or not the non-randomized trials give any sort of bias in the statistical tests, is reduced to the following:

(1) Whether or not, in cases when the hypothesis tested H_0 is true, and, in particular, when $\bar{A} = \bar{B}$, the value of $t = \bar{x}/(s/\sqrt{n})$ calculated by this or that method exceeds the fixed value of t_α with the frequency $\alpha = P/2$ prescribed by the theory.

(2) Whether or not, in cases when the hypothesis H_0 is wrong and thus $\bar{B} - \bar{A} = \Delta > 0$, the t test detects this circumstance, the value of t falling above the critical t_α, with a frequency predicted by the theory.

If, on any empirical evidence, either of the above two questions were to be answered in the negative, then we should say that the mathematical model

that served as a basis for calculating $t = \bar{x}/(s/\sqrt{n})$ does not correspond to the actual trials, and that either the model or the experimental design should be altered. If, however, a considerable volume of empirical data fails to deny either 1 or 2, then the practical man would probably say that, from a purely academic point of view (which may be interesting by itself), there may be disagreements between the experimental technique and its mathematical model, but that these disagreements do not concern him. In fact, the statistical test gives all it is expected to give; it rejects the hypothesis tested H_0 when it is in fact true as frequently as expected, and it detects the falsehood of H_0 when it is wrong with about the same frequency as predicted by theory.

It is seen, therefore, that the whole question is reduced to what is the actual empirical distribution of values of t in cases when $\bar{A} = \bar{B}$, and in cases when $\bar{B} - \bar{A} = \Delta > 0$. We must discuss the question of how such empirical distributions can be obtained.

It is easier to obtain an empirical distribution of t for the case when $\bar{A} = \bar{B}$ than for the case $\bar{B} - \bar{A} > 0$. We have to use for this purpose the results of so-called *uniformity trials*. Imagine a large field divided into a number of very small plots, considerably smaller than the ones used for actual experiments. To avoid misunderstanding, we shall call them elementary plots. If you treat all these plots in exactly the same way, so far as possible, and sow them with the same variety, you will have a uniformity trial. The results of such trials, represented by a plan of the experimental field with the yields of single elementary plots, are to be found in various publications. However, not all of them are equally suitable for our purpose, mainly because the elementary plots used are not sufficiently small, or because they differ considerably from squares. If the elementary plots are very tiny squares, then they can be combined in various ways to form what could be real experimental plots. If we wish to see what the results of some particular experiment on this field would be, as in comparing some objects A, B, $\cdots$, which *are* in fact identical (though we are not aware of it), we simply assign these hypothetical objects to particular plots and then perform all the calculations on the figures provided by the uniformity trial and apply the tests that we should apply if we had to deal with an actual experiment. If the elementary plots are large or very long, then the same procedure can be applied; but it may be hard to produce experimental plots of the desired size and shape.

For our purpose we should need uniformity trials with elementary plots that could be combined into half drill strips. Suppose that many such hypothetical half drill strips are available in the form of a table like the following, where each rectangle represents a half drill strip and the figure written on it the sum of the yields of the elementary plots of the uniformity trial of which the experimental plot is composed. They would be the actual

101	107	102	97	101	102	106	113	114	106	99	101
↑	↑	↓	↓	↑	↑	↓	↓	↑	↑	↓	↓
A	B	B	A	A	B	B	A	A	B	B	A

etc.

yields obtained on these plots in an experiment with two hypothetical but identical varieties A and B. Writing in successive letters A, B, B, A, etc., on the plan of the hypothetical experiment (as shown), and applying any given mathematical model, we can calculate t, knowing that it refers to the case where $\bar{A} = \bar{B}$. A set of such values of t, calculated from the results of a number of uniformity trials, will produce the distribution we want to compare with the theoretical one deduced by Student, namely,

$$p(t) = C(1 + z^2)^{-n/2}, \tag{7}$$

where $t^2 = z^2(n - 1)$, and $n - 1$ is the number of degrees of freedom on which the estimate s^2 is based.

If the sandwiches are randomized, then the estimate of $\bar{B} - \bar{A}$ is simply the arithmetic mean $\bar{x}$ of the numbers x_i as defined above, and

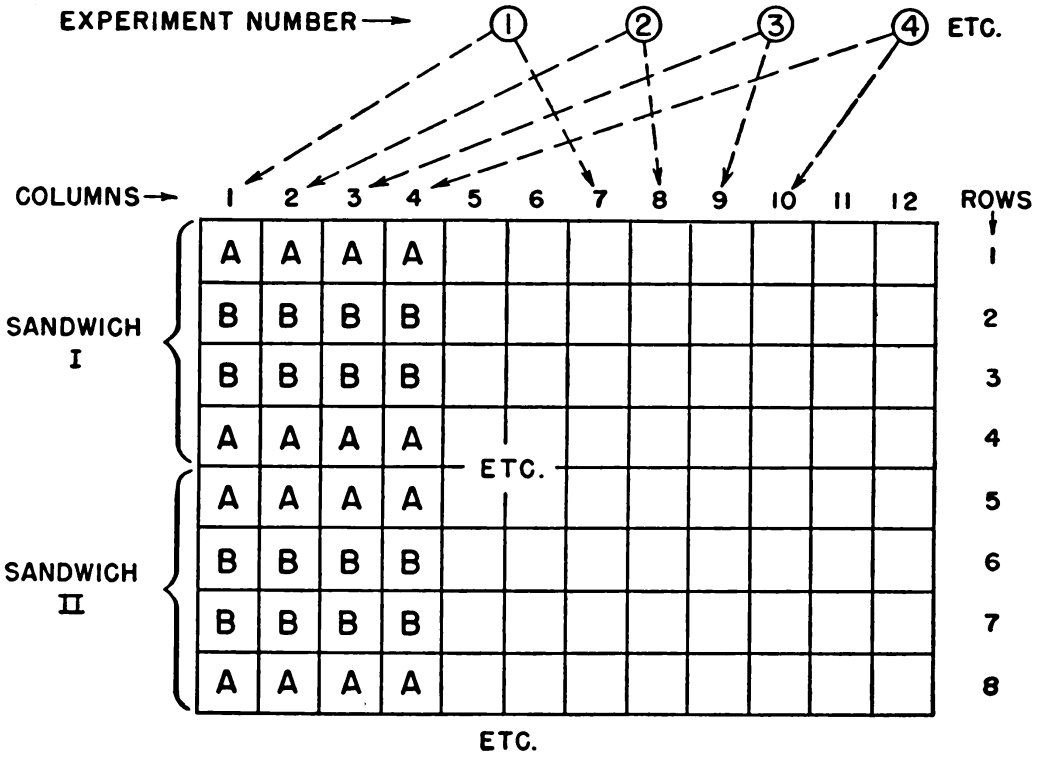
$$\frac{s^2}{n} = \Sigma \frac{(x_i - \bar{x})^2}{n(n - 1)}. \tag{8}$$

As far as I am aware, the first authors to run tests on uniformity trial data to see whether or not the distribution of $\bar{x}/(s/\sqrt{n})$ from *non-randomized* sandwiches followed Student's frequency of t, were S. Barbacki and R. A. Fisher.[6] They came to the conclusion that the lack of randomization is destructive to the t test, and they blamed Student for thinking differently. It seems to me, however, that Barbacki and Fisher were a little unfair to Student, and that the figures they produced are entirely valueless.

Barbacki and Fisher took just one uniformity trial for which weights of yields of wheat on short parts of single rows were published.[7] They combined the adjoining rows to obtain the width of a half drill strip. The rows were long and they divided them into 12 columns and so obtained 12 columns of hypothetical half drill strips, each being a continuation of the strips in other columns. These columns were interpreted as representing the results of six hypothetical experiments comparing some variety A

[6] S. Barbacki and R. A. Fisher: "A test of the supposed precision of systematic arrangements." *Annals of Eugenics,* Vol. 7 (1936), pp. 189–193.

[7] G. A. Wiebe: "Variation and correlation in grain yields among 1500 wheat nursery plots." *J. Agric. Res.,* Vol. 50 (1935), pp. 331–357.



with another B. Experiment No. 1 would consist of sandwiches in columns 1 and 7; experiment No. 2 would consist of sandwiches in columns 2 and 8; etc., as marked in the figure. The two authors calculated t for each such experiment and were pleased to find that, in spite of the fact that the hypothetical varieties A and B were identical, the distribution of the empirical t was far from similar to the theoretical one. In fact, all values of t had the same sign! This, of course, was to be expected because the values thus calculated were not independent. It is known that the direction of rows is frequently that of ploughing and that in this direction we frequently observe what I call *waves of fertility:* if one of the plots in the first row is better than the corresponding plot in the second, then this is likely to be true for all other plots in these rows. These waves of fertility are very marked on the field used by Barbacki and Fisher and consequently the value of t calculated for any one of these hypothetical experiments could not be much different from the one for any of the others. The whole argument is as if we would toss a penny just once, look at it six times and, having recorded six heads, argue that the penny must be biased. The authors are unfair to Student because he called attention to the fact that parts of the same strip are highly correlated.[8]

[8] Student: "On testing varieties of cereals." *Biometrika,* Vol. 15 (1923), pp. 271–293. See pp. 286–287 in particular.

It follows that we can not accept the results of Barbacki and Fisher as conclusive in the question which interests us. Their figures emphasize only the known fact that there is danger in replicating an arrangement on plots in adjoining columns because an error in one of the columns is likely to be repeated in the others. This does represent an advantage for the randomized arrangements but does not show that systematic experiments, if carried out with due precautions, necessarily give biased results.

There is no doubt, however, that the application of the formula (8) does represent a crude treatment. This was recognized by Student who, in a paper published in the *Supplement to the Journal of the Royal Statistical Society,* Vol. III, pp. 114–136, 1936, suggested a new way of proceeding. This is based on the hypothesis that the level of fertility along the row of drill strips is either rising or falling off more or less regularly, so that, within each pair of half drill strips, the fertility of the next half drill strip differs from that of the preceding one by a fixed quantity, which Student called the *linear fertility slope.* Again, there is no doubt that this assumption does not correspond exactly to what happens in practice, but the formulas that the new mathematical model involves—let it be called the new Student's method—have a greater chance of giving satisfactory results than formula (8). In fact, this method along with that of parabolic curves, is based exclusively on the assumption that the experiment is arranged systematically. Whether or not it works well must be tested empirically.

Some work designed to throw light on the question in which we are interested has been done by one of my students, Mr. C. Chandra Sekar. He tried to collect as many uniformity trial data as he could possibly find, and on each field he arranged a number of independent hypothetical experiments in systematic half drill strips. The total number of experiments was 120. For each experiment the value of t was calculated twice, first by the new Student's method and then by the method of parabolic curves. The distributions obtained are shown in Figures 1 and 2. In each case the empirical distribution was compared with the theoretical Student's distribution using the smooth test [9] for goodness of fit. The symbol $P\{\psi^2 \geqq \psi_0^2\}$ represents the probability of obtaining by chance an agreement between theory and observation worse than that actually observed. For the new Student's method this probability is .173 and for the method of parabolic curves, .643. The two graphs and the two probabilities represent the empirical part of the inquiry. Whether the agreement between the theory and the observation is or is not satisfactory is a subjective question. However, I submit that, especially as regards the method of parabolic curves, one could hardly expect anything better.

[9] J. Neyman: "'Smooth test' for goodness of fit." *Skandinavisk Aktuarietidskrift,* Vol. 20 (1937), pp. 149–199.

FIGURE 1

t distribution in the half-drill-strip experiments

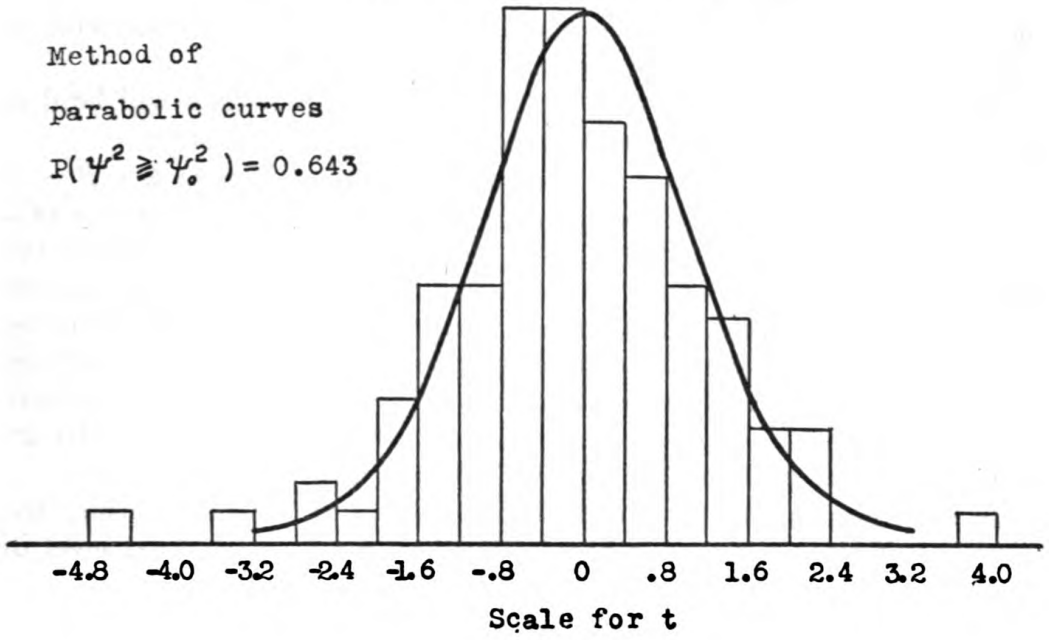
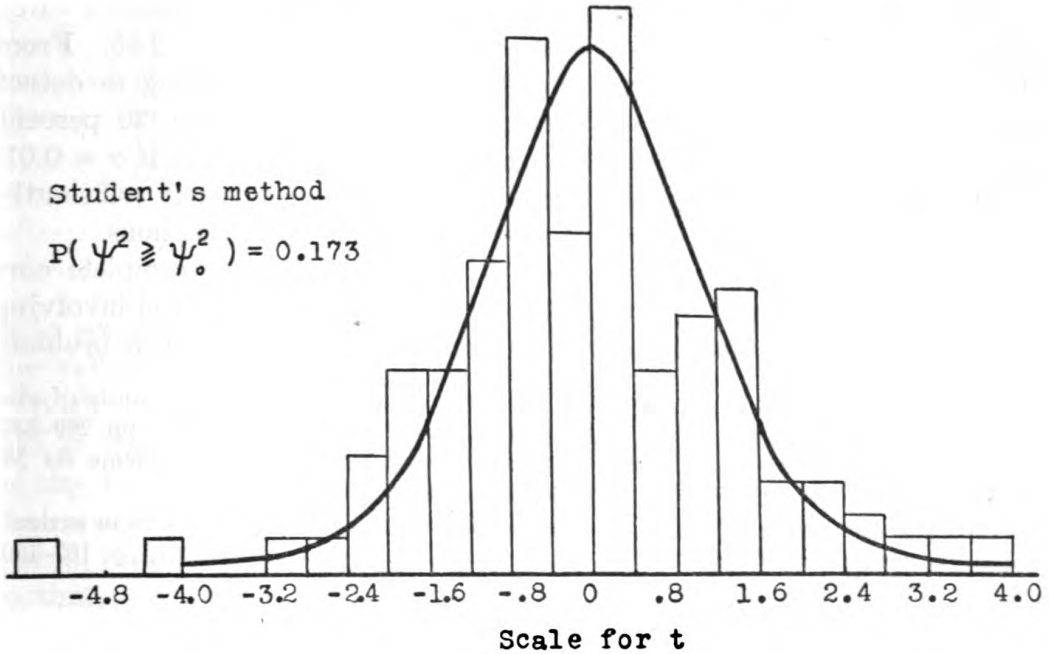


FIGURE 2

t distribution in the half-drill-strip experiments



Now let us turn to the question of the effectiveness of the two methods in cases where one of the varieties, say B, is actually better than the other, A. In relation to this situation and on the assumption that the observations are mutually independent and follow the normal distribution, the theory of the t test is as follows.

(i) It has been shown [10] that the superiority of B over A will be discovered by the t test more frequently than by any other test.

(ii) The frequency of the t test failing to detect a difference $\Delta = \bar{B} - \bar{A}$ when it actually exists and is equal to ρ times the *true standard* error σ of $\bar{x}$ is known and depends on the number of degrees of freedom on which the estimate of σ is based. This is what is technically called the probability of an error of the second kind. The first short table of this kind was published by S. Kolodziejczyk.[11] This was later supplemented in a joint paper by K. Iwaszkiewicz, S. Kolodziejczyk, and myself,[12] wherein certain graphs are published, two of which are shown on pages 79–80. Finally, a differently arranged table was published by Miss B. Tokarska and myself.[13]

In these graphs n means the number of degrees of freedom on which the estimate of error variance is based. Further, α means the fixed level of significance with which you work. To make the diagrams clear let us consider an example. Suppose you are arranging a randomized blocks experiment with six treatments and three replications. In this case $n = 10$. From previous experience you know that the standard error per plot is likely to be, say, 10 percent of the average yield, and you want to know the probability that the experiment will fail to detect as large a difference between your treatments as 20% of the general mean. The expected value of your σ is $10\sqrt{2/3} = 8.16$. Your $\Delta = 20$, and $\rho = 20/8.16 = 2.45$. From the diagram you find that the probability of the t test failing to detect the difference between the treatments when it is as large as 20 percent of the average yield is about 0.25 if $\alpha = 0.05$, and about 0.55 if $\alpha = 0.01$. You will probably decide that the experiment planned is not sufficiently accurate, and you will try to increase the number of replications.

Of course, points (i) and (ii) refer to the ideal case of a complete correspondence between the experiments and the mathematical model involving the normal distribution and mutual independence of "errors." Our problem

[10] J. Neyman and E. S. Pearson: "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Royal Society, London,* Vol. 231-A (1933), pp. 289–337.

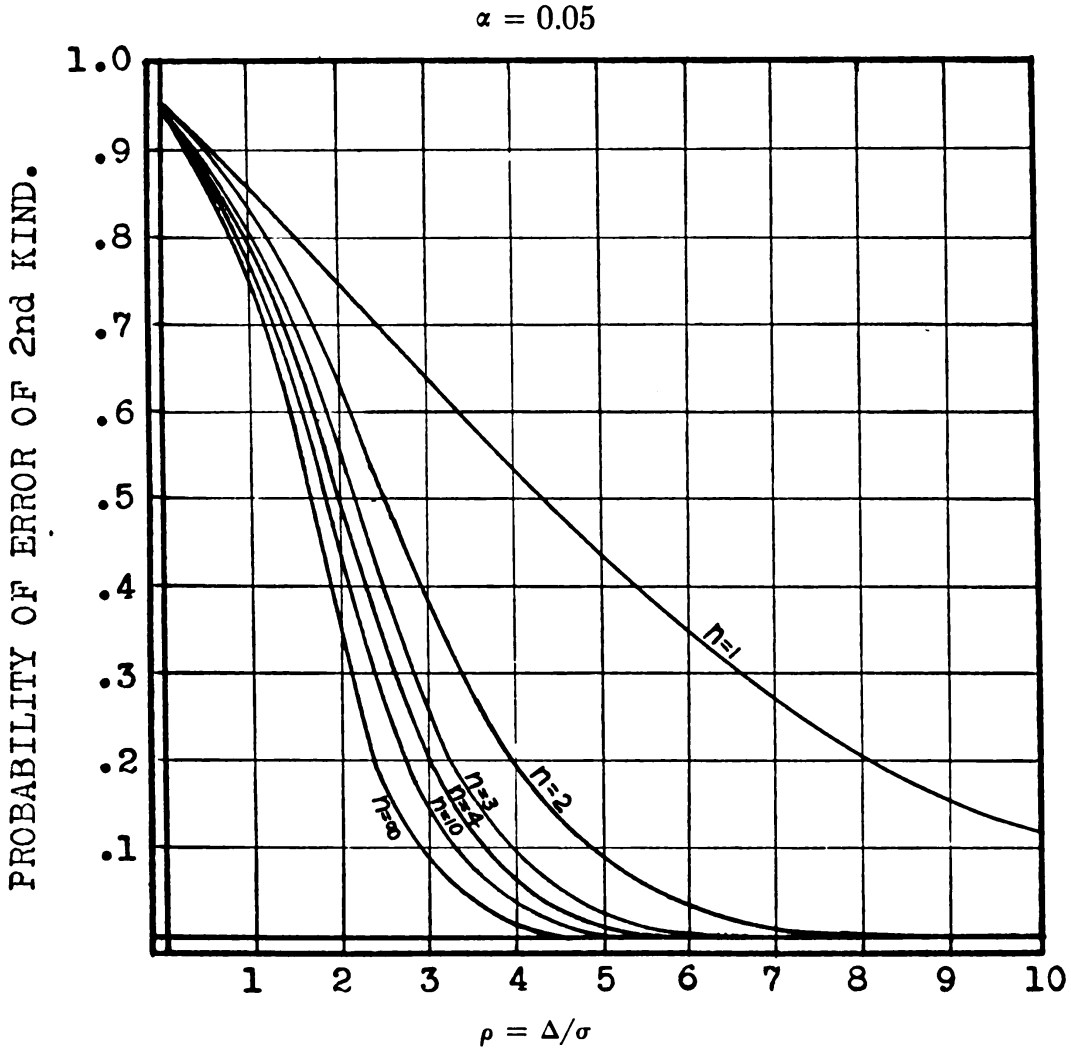
[11] S. Kolodziejczyk: "Sur l'erreur de la seconde catégorie dans le problème de M. Student." *Comptes Rendus,* Vol. 197 (1933), pp. 814–816.

[12] K. Iwaszkiewicz, S. Kolodziejczyk and J. Neyman: "Statistical problems in agricultural experimentation." *Supplement to Jr. Roy. Stat. Soc.,* Vol. 2 (1935), pp. 107–180. See pp. 133–134 in particular.

[13] J. Neyman and B. Tokarska: "Errors of the second kind in testing 'Student's' hypothesis." *Jr. Am. Stat. Assoc.,* Vol. 31 (1936), pp. 318–326.

FIGURE 3

Diagram showing dependence of probabilities of second kind errors on ρ and n, when $\alpha = 0.05$

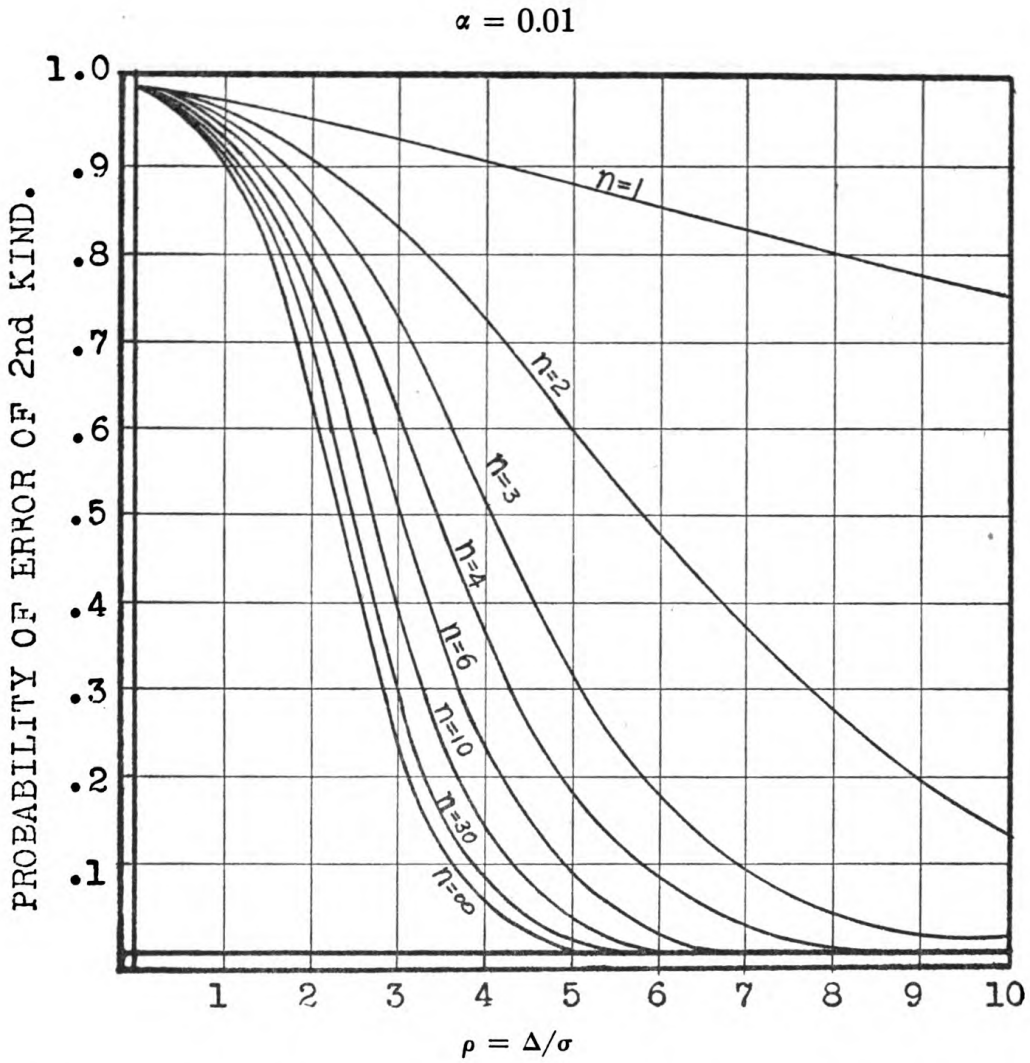


is to see whether or not the existing divergences from this model influence the validity of the theoretical conclusions.

With regard to point (i) raised above, there are insurmountable difficulties in this respect. There is no way to produce *empirical* evidence that in any fixed conditions of experimentation it is impossible to invent a test that would be more sensitive than the t test. If any other test were suggested, then we could produce empirical results comparing its sensitiveness to that of t, and this comparison might show that the alternative test is better than t. But any number of such comparisons, all of them favorable

FIGURE 4

Diagram showing dependence of probabilities of second kind errors on ρ and n, when $\alpha = 0.01$



to t, would not prove that the t test is actually the best. For this reason, and because no test alternative to t has been suggested, we shall drop the question of an empirical test of question (i).

An empirical test of point (ii) is much easier, though it requires a lot of calculations. In fact, the problem is very similar to that dealt with in the case where A was identical with B. We start by producing what could be the results of actual trials in half drill strips, including the actual inequalities in soil fertility and the actual experimental errors, in which, however, the true average yield of B is greater by a certain amount than

that of A. For each such experiment we calculate the value of t and see how frequently it fails to exceed the critical tabled value of t, that is to say, how frequently the t test fails to detect the advantage of B over A. This frequency must then be compared with the probability of an error of the second kind to be found in the tables mentioned above or read from the graphs on pages 79–80.

In order to produce the quasi-empirical data for the above purpose we use again the same uniformity trials that were used before. I have mentioned on page 73 that on each of the fields with uniformity trials it is possible to arrange more than one hypothetical experiment in half drill strips. Each of them gives an estimate of the error variance. Several such estimates were averaged, and this average was taken as the true value of the error variance for the experiments on any particular field.

To see more clearly what was done next, consider the situation on any two particular fields. The assumed true standard deviations of the estimates of $\bar{B} - \bar{A}$ on those fields are σ_1 and σ_2, respectively. Using the graphs of probabilities on pages 79–80, the values $\rho(20)$, $\rho(40)$, $\rho(60)$, and $\rho(80)$ of ρ were found, for which the probabilities of errors of the second kind are 0.20, 0.40, 0.60, and 0.80. These values of ρ were than multiplied by σ_1 and σ_2 to obtain what I shall denote by $\Delta_1(20)$, $\Delta_2(20)$, $\Delta_1(40)$, etc., so that, for example,

$$\Delta_1(20) = \sigma_1\rho(20), \qquad \Delta_2(20) = \sigma_2\rho(20), \text{ etc.}$$

You will notice that $\Delta_1(20)$ represents the value such that if the difference between $\bar{B}$ and $\bar{A}$ tested on the first field were equal to $\Delta_1(20)$, then the theoretical probability of the t test failing to detect the advantage of B over A would be exactly equal to 0.20.

Suppose that the values of $\Delta_i(20)$, $\Delta_i(40)$, $\Delta_i(60)$, and $\Delta_i(80)$ are calculated for the ith field. Take one of the hypothetical experiments in the systematic half drill strips previously arranged on some particular field from data of uniformity trials, and add $\Delta_i(20)$ to all the hypothetical yields of the object B. Before this addition, the variability of yields from plot to plot was due solely to soil variation and technical errors, since all the plots were equally treated and sown with the same variety. After the addition of $\Delta_i(20)$ to the yield of the hypothetical B, we obtain what could be the result of an actual trial of A and B, including the effect of soil variation and technical errors, $A - B$ having the property that whatever the true yield of A, the true yield of B is greater by the amount $\Delta_i(20)$. That is what we want for testing the distribution of t when $\bar{B} - \bar{A} = \Delta_i(20)$.

Mr. C. Chandra Sekar calculated t for each of the experiments in such systematic sandwiches, obtained in the above way from the data of uniformity trials. Again, both the new Student's method and the method of parabolic curves were tried. The results, in the form of frequencies of

non-detection of the advantage of B over A, both observed and theoretical, are set up in the following table.

TABLE I

Relative frequencies of failure to detect a real advantage of B over A in systematic half-drill-strip experiments

Theory, percent	Method of parabolic curves, percent	Student's method, percent
20	23.3	27.5
40	40.8	46.7
60	62.5	61.7
80	78.3	75.8

Again, this is the objective part of the answer to the question of whether or not the lack of randomization ruins the t test. The first column gives the theoretical frequency of cases in which the t test should fail to detect the advantage of B over A. The other columns show what these frequencies would be in a number of experiments in which the variability of the soil and the experimental errors are exactly as they were in actual uniformity trials. Is the disagreement sufficient to say that the t test is of no use when applied to the systematic half drill strips? This, as I said, is a personal question. So far as I am concerned, the agreement between the theory and the empirical results seems to be satisfactory. Especially in the case of parabolic curves, the t test both detects the advantage of B when this advantage exists and suggests its existence when it does not exist with relative frequencies very much the same as indicated by the theory.

In consequence, I do not see any evidence to support the assertion that lack of randomization by itself is ruinous to statistical tests. We must, however, remember the following points.

(i) The above empirical results refer to one particular systematic arrangement in half drill strips: $ABBA$, etc. It is reasonable that if we take any other systematic arrangement, the conclusions suggested by the empirical results may be different. If we take the systematic arrangement of blocks with more than two objects

ABCA, ABCD, $\cdots$,

then probably the advantage of the method of parabolic curves over the ordinary formulas for randomized blocks will be more marked than in the case of half drill strips, but this requires an empirical test.

(ii) The waves of fertility are an important feature that should be borne in mind in any case and especially when the trials are arranged systematically. Whenever I was able to ascertain the direction of ploughing, I found that the fertility seems to stay steadier along the direction of ploughing than across. It seems to me that the direction of ploughing may be the real cause of these waves, but I have no definite evidence of this. Sometimes the waves are difficult to detect when you simply look at the uniformity trial data. In other instances they are very pronounced. The following table gives a part of the uniformity trial data with rye as described by Hansen.[14] Looking at it you will hardly believe that all the plots were sown with the same variety and equally treated, but this is a fact.

TABLE II

Hansen. Yields of rye. Uniformity trial data, 1909

Probable direction of ploughing ↑↓

1	2	3	4	5
101	84	113	88	110
107	91	114	88	109
102	94	106	84	106
97	94	99	88	105
101	90	101	84	104
102	86	99	84	102
106	90	100	85	104
106	92	104	85	105

Imagine now that, without knowing the peculiar fertility level of the field, you use this field for an actual experiment and cut your plots along the columns. The results would be deplorable. On the other hand, if long and narrow plots were cut *across* the columns, the experiment might have been fairly successful.

If practical circumstances forced one to cut the plots along the columns of the above, say four rows deep, so that out of each column we had two plots, then it would be most inadvisable to arrange a systematic experiment replicated exactly in the two rows, e.g.,

[14] N. A. Hansen: "Prøvedyrkning paa Forsøgsstationen ved Aarslev." *Tidsskrift for Planteavl,* Vol. 21 (1914), pp. 553–617.

ABCD, ABCD, ···

ABCD, ABCD, ···

since the second row would repeat almost identically the same soil errors as there are in the first. In such circumstances, a randomized arrangement would be most useful. In this sense, the randomized arrangements do have definite advantages over the systematic ones.

Turning to the question of the waves of fertility, I think that from the point of view of accuracy of agricultural trials it would be most useful to have some indication of their cause. Probably it would not be too difficult to make a special experiment to discover whether the direction of the waves of fertility is actually connected with that of ploughing.

Part 2. On Certain Problems of Plant Breeding

(The contents of this lecture are based on a conference held in Room 4090 of the Department of Agriculture, April 7, 1937, 10 A.M., Dr. S. C. Salmon presiding.)

The problem I am going to discuss in this conference is a specific one connected with the breeding of new varieties of sugar beet. However, I believe that it is of wider interest than its restricted nature would indicate. Aside from the fact that similar problems arise in breeding other plants, there is another and a stronger reason for my choice of this particular subject. The point I want to illustrate is this: the methods of mathematical statistics may be useful not only in treating isolated trials as, for example, those discussed in the preceding conference but also in forming the over-all policy of an organization. The particular organization about which I will speak is a sugar beet breeding establishment, but it can be seen that problems of a similar kind will arise elsewhere.

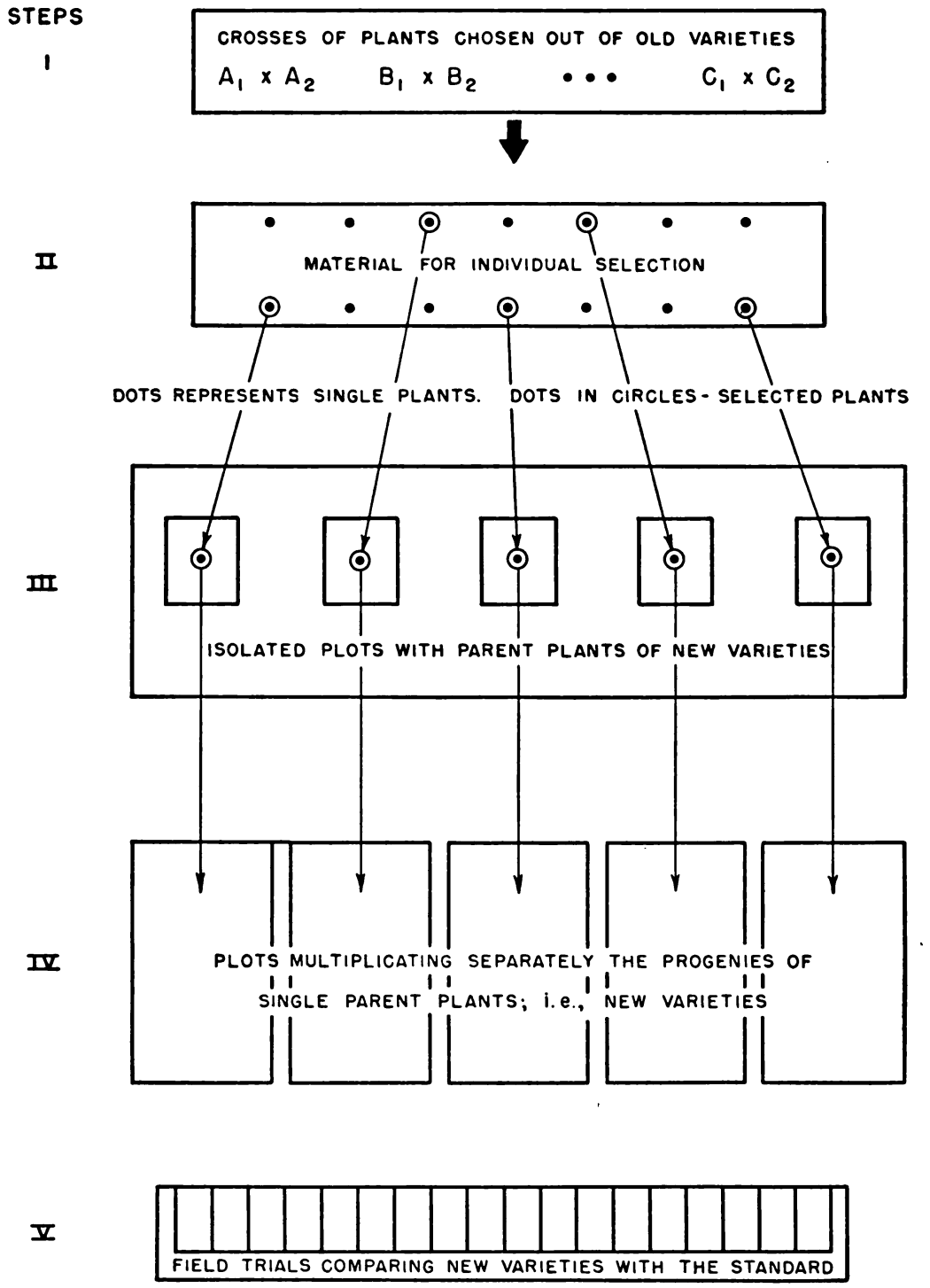
The idea of the problem originated from contact with sugar beet breeders in Poland. However, the results that I am going to present are due to Mrs. Y. Tang, M. Sc., and all of the details are published in her paper prepared at the Department of Statistics, University College, London.[1]

The process of breeding new varieties of sugar beets is fairly complicated, but a rough idea of its essence can be obtained from the diagram on page 85 which represents schematically five distinct steps. In considering these steps, we must remember several important points concerning sugar beets. The first is that the sugar beet is a two year plant. During the first vegetative season a seedling produces a plant with a big root containing a considerable amount of sugar but yielding no seeds. The seeds are produced in the course of a second vegetative season when the

[1] Y. Tang: "Certain statistical problems arising in plant breeding." *Biometrika,* Vol. 30 (1938), pp. 29–56.

FIGURE 1

Plant breeding: scheme of production of new varieties of sugar beets



plant uses the food previously accumulated in its roots in the form of sugar. The second important point consists in the fact that the sugar beet is a cross fertilizing plant, and that this makes it extremely difficult, if not impossible, to produce anything like a pure line. Finally we must remember that we may have various aims in the production of new varieties: we may try to produce (i) beets with highest sugar content, (ii) beets with the highest yield of roots per acre, or (iii) beets with the highest yield of sugar per acre. The discussion which follows applies to all three cases, but we shall consider only the first.

Keeping these points in mind, let us consider the diagram and see what are the five consecutive steps leading to new varieties. The first step consists in choosing from the existing varieties a number of roots which, for various reasons, seem to be promising, and in forcing them to cross fertilize. For this purpose the roots are planted in pairs on plots isolated from one another in a larger field of some cereal. The hope is that the capacity of producing high sugar content in old varieties may be increased as a result of crosses between them. But it is clear that a cross must sometimes increase the capacity of producing a low sugar content. Therefore not all of the progeny of the crosses are suitable for further breeding, and we have to perform a selection.

All the seeds produced by the crosses are sown on a larger plot and produce roots. These form the material for what is called "individual selection," the second step in our scheme. At the end of the vegetative period all the roots are lifted, washed, and weighed. A small portion is cut from each root and analyzed for sugar content. This cutting neither kills the root nor affects its ability to produce seeds as well as if it had been left intact. The majority of roots analyzed are discarded as unsatisfactory. The remaining ones, having the highest sugar content or certain morphological characteristics indicating that they may be able to produce high sugar content, are stored for the winter. Then, in the spring they are planted separately on isolated plots to produce seeds, mostly from self-fertilization. This is the third step in our scheme. Each of the selected roots is called a parent plant, and originates a new variety.

Obviously each parent plant is able to produce only a very limited amount of seed. Therefore, two or more vegetative seasons must be used to multiply the seeds of the new varieties, and this is described in the diagram as step IV.

The fifth and last step consists in determining which of the newly bred varieties possess an advantage in sugar content over some established standard. We must remember that the sugar content of any individual root depends not only on the genetical composition of the plant but also, frequently to a greater extent, on various conditions of environment. Conse-

quently the sweetest of the parent plants selected in step II do not necessarily produce the varieties with the highest sugar content. Also it is possible that still sweeter varieties might have been produced by some of the roots grown in step II that, owing to uncontrollable variation of environment, had small sugar content and were discarded. The field trials (step V) are meant to eliminate the individual variability of sugar content in roots of a new variety. We may put it otherwise: analyses in V are a comparison of varieties, wherein the properties of individual roots are more or less ignored.

Needless to say, along with the field trials in step V we continue to multiply the seeds of the new varieties, and the final decision as to whether or not any one of them is a success is made, not after one year, but after several years' trials. However, these are details.

In any event, after the fifth step is concluded, the breeder has to decide which of the new varieties are suitable to put on the market. Other families of beets are discarded as failures.

I must call your attention to certain consequences of the fact that the sugar beet is a cross fertilizing plant and consequently that any single individual is heterozygous with respect to a number of pairs of genes. One consequence is that a plant which is called a "new variety" does not represent anything stable, but changes from generation to generation.

Further, according to a law discovered by Galton and which is a consequence of the Mendelian laws, the change is unfavorable to the breeder: there is necessarily a regression (i.e., a set-back) in sugar content. This makes it impossible for the breeder to find just one or two exceedingly sweet varieties and keep them for reproduction from year to year without further selection. After a relatively short period, the sugar content of new generations will drop and the breeder will lose his market. Consequently, each breeder has to repeat constantly the steps described above, perhaps with certain modifications, and to start step I each year, meanwhile continuing the following steps applied to varieties planted in previous years.

Another consequence of the instability of the varieties is the instability of the standard variety, with which the new varieties are compared in step V. As each variety changes necessarily from year to year, so must the standard change, even if it bears the same label.

In Poland it is usual to take as standard that variety which in the preceding year proved to be the sweetest. The beet sugar industry arranges each year competitive experiments with a number of varieties, produced by several leading firms. These experiments, carried out in a number of places in all the beet growing districts of Poland, are made according to a certain fixed method, with the same number of replications, etc.

After this somewhat lengthy preliminary, we may turn to the problems which the breeder must face in deciding on details of his work. These problems are statistical in character and refer to steps II and V. Their aim is to see how the breeder is likely to increase his chances of success. We must now review some of the possible causes of his being unsuccessful.

1. The breeder may be unlucky in choosing plants for his crosses in I. But this is not a statistical problem.

2. Supposing that the breeder was successful in I, he may be unlucky in II by failing to select for further breeding the roots that have the best genetical properties. This is a problem that is partly botanical and partly statistical. The statistician may advise the breeder to *select for further breeding as many parent plants as he possibly can, so as not to omit the best ones.* I shall call this *advice A.*

3. Suppose now that the breeder was successful both in steps I and in II, and, consequently, that *some* of his new varieties that come for comparison with the standard in V are better than the standard. Obviously, again he may be unlucky and lose these new varieties. The accuracy of field trials is known to be limited and it is just possible that through unavoidable errors the experiments will fail to detect the goodness of the best varieties, so that eventually they will be discarded. This, of course, would be most unfortunate, since it would mean a total waste of a considerable amount of effort, money, and time. Here again is a problem for the statistician and he will give what I shall call *advice B: make your experiments as accurate as possible; if you cannot improve the method of experimentation, then increase the number of replications.*

Both advice A and B are sound, of course, but both will seem very troublesome to the practical breeder. His means are always more or less limited and, before all, this applies to the arable area at his disposal. You will notice that each of the advices A and B makes a claim on this area and the breeder is faced with the dilemma: to select more roots in step II and then make fewer replications in the comparative trials in step V, or to select fewer roots, to start fewer varieties each year, and then to compare them with the standard, using many replications. If he selects too few roots in step II, he is likely to have poor material from which to choose and he may be unsuccessful even though his trials in step V are very accurate. If he starts a great many new varieties in step II, his chances of having some good ones are high, but if the trials in step V have few replications, the best new varieties may go undetected.

The decision as to the number of new varieties to be started each year and as to the number of replications in the comparative trials to be made is just the matter of the breeder's general policy which I want to discuss. This is not a problem strictly limited to plant breeding. Its generality may

be judged from the following example taken out of an entirely different field. Imagine a squadron of twelve bombers facing twelve ships of an invader. Each of the twelve bombers may be directed to attack a separate ship. The argument is that, given good luck, all of the twelve ships may be sunk. On the other hand, the twelve bombers may be directed to attack, say, three selected ships, four planes to a ship. In this case, no more than three ships can be sunk, but it is obvious that the chances of sinking at least one ship are much better. The problem of the right distribution of the attacking air force among available targets of equal priority is essentially the same problem of general policy which is faced by the plant breeder.

The policy problem of the plant breeder was dealt with by Mrs. Tang with particular reference to sugar beets. Her results show how to calculate approximately the results of plant breeding for any given ratio of the number of new varieties and the number of replications used. Of course, the final appreciation of the results of such calculations must depend on many local conditions.

It is interesting to note that the solutions of the above problem, advanced by practical breeders, most probably on intuitive grounds, differ enormously. The number of new families of sugar beets started yearly by Polish breeders goes into hundreds, while the number of replications they use is sometimes as small as four, and to my knowledge, has never exceeded sixteen. On the other hand, the breeders of barley in England and Ireland start with only four or perhaps five new families and then test them in 40 half drill strips! It is entirely possible that this difference is due to special characteristics of the two particular plants and also to the cost of land, labor, etc. But it is possible also that the general intuition of the practical worker was, in one case or in the other, misled.

Now I must recall the nature of the errors that may be committed when testing statistical hypotheses. In doing so, I will treat the particular case of the comparison between a new variety V and the standard S. Denote by $\overline{V}$ and $\overline{S}$ the true average sugar content that the two varieties would yield if each were sown on the entire experimental field and if there were no technical errors. We are interested in the difference

$$\Delta = \overline{V} - \overline{S}, \tag{1}$$

which may be termed the true sugar excess of the variety V over the standard or, for short, the *sugar excess*. If Δ is positive, the new variety will be considered satisfactory. Otherwise it will be a failure. The experiment does not give us the true value of Δ but only the estimate x of Δ which is always affected by a positive or negative experimental error ϵ, so that

$$x = \Delta + \epsilon. \tag{2}$$

Before the variety V is placed on the market, the breeder wants to have some "evidence" that it is satisfactory, i.e. that Δ (not x) is positive. He must be particular on this point for frequently, otherwise, he will have inferior goods and lose his customers. In this instance, mathematical statistics is helpful and provides means by which the frequency of cases when Δ is judged positive without, in actual fact, being positive can be reduced to any low level α chosen in advance. α is called the level of significance.

Statistically, the problem of the breeder is reduced to testing the hypothesis H_0 that

$$\overline{V} - \overline{S} = \Delta \leqq 0. \tag{3}$$

If, as a result of our test, we decide to reject the hypothesis H_0, this is equivalent to a recognition that we have "evidence" of Δ being positive, i.e. of the new variety being better than the standard.

The test of the hypothesis H_0 consists in the rule of rejecting H_0 whenever

$$\frac{x}{s} > t_\alpha, \tag{4}$$

where s is the estimate of the standard error of x and t_α is a constant number taken from Fisher's tables corresponding to the number of degrees of freedom on which the estimate s is based and to the tabled $P = 2\alpha$. This test was originated by Student.

The properties of this test are: (i) whenever the new variety is barely as good as the standard, i.e. when $\Delta = 0$, the hypothesis tested will be rejected (this is equivalent to placing an unsatisfactory variety on the market) with a relative frequency equal to α; (ii) whenever H_0 is true and the new variety is worse than the standard, i.e. when $\Delta < 0$, the relative frequency of rejection will be even smaller than α; (iii) whenever H_0 is wrong and the new variety is superior to the standard, i.e. when $\Delta > 0$, then the above test will detect this circumstance more frequently than any other imaginable test having properties (i) and (ii).[2]

We must be clear on this point and, therefore, let us consider some numerical illustrations. One breeder A may desire that the proportion of his unsuccessfully bred varieties which reach the market should not exceed 5 percent. In this case, the level of significance being $\alpha = 0.05$, he finds in Fisher's Table IV the value of t corresponding to $P = 2\alpha = 0.10$. If the number of degrees of freedom is 12, then $t = 1.782$. Thus he will reject the hypothesis H_0 and say that his variety is good enough to be put on the market when $x > 1.782s$. Another breeder B may consider that to allow 5 percent of his unsatisfactory varieties to go on the market

[2] J. Neyman and E. S. Pearson: "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc.*, London, Vol. 231-A (1933), pp. 289–337.

is too great a risk; he may consider that the proportion of such varieties should not exceed 1 percent. In such a case, he would put $\alpha = 0.01$ and select t corresponding to $P = 2\alpha = 0.02$. On this basis he would let through his new variety only if $x > 2.681s$. Other breeders may be even more cautious.

QUESTION BY DR. SARLE: Is there any danger of being too cautious?

ANSWER: Yes, there is, and I am most grateful for the question. The danger consists in that, whenever we are too particular in trying to avoid unjust rejections of the hypothesis tested, i.e. rejection when it is in fact true, then we are exposing ourselves to an increased risk of failing to detect cases when the hypothesis is false. This problem is sufficiently important to justify a little digression.

It will be convenient to use the special terminology introduced in Chapter I, Part 3, to distinguish between the two kinds of error that we may make when testing a statistical hypothesis and, in particular, when judging whether a given variety is or is not better than the standard. If, as a result of a test, we reject a hypothesis when in fact it is true, we say that the error committed is of the first kind. Thus, when the breeder puts on the market a variety that does not exceed the standard, he commits an error of the first kind. On the other hand, an error of the second kind consists in accepting the hypothesis tested when in fact it is false. Thus, when the breeder does not find sufficient reason for judging his variety satisfactory (i.e. when $x/s \leqq t_\alpha$), whereas his new variety *is* actually sweeter than the standard (i.e. $\Delta > 0$, though he does not know it), he commits an error of the second kind.

Errors of the first kind are dangerous to the trade of the breeder, but then so are errors of the second kind. It must be remembered that each rejection of a satisfactory variety means a complete waste of effort and money spent for a substantial number of years: after all the years of work a variety exceeding the standard in sugar content is successfully produced and then an error of the second kind causes this variety to be discarded. Thus it is necessary to have as clear an idea as possible regarding the chance of committing an error of the second kind. Numerical evaluation of the probabilities of errors of the second kind are based on charts reproduced in the preceding chapter.

In the present notation, the "standardized" error of the second kind is

$$\rho = \frac{\Delta}{\sigma} = \frac{(\overline{V} - \overline{S})}{\sigma}. \tag{5}$$

This is the true value of Δ divided by the true value of σ, where σ is the true standard error of x (not the estimate s of σ).

To illustrate the use of the diagrams in answering the question raised by Dr. Sarle, we suppose that the arrangement contemplated for a future

experiment is in randomized blocks with three varieties and six replications which gives $n = 10$ degrees of freedom. We suppose further that previous experience indicates that σ may be taken as something like 0.5. In these circumstances, let us see what would be the chance of detecting that a particular variety is better than the standard when Δ is actually positive and as large as 1 percent. In order to answer this question, we calculate $\rho = \Delta/\sigma = 2$ and refer to the curves corresponding to $n = 10$ on pages 79–80. We see that if we use the level of significance $\alpha = 0.05$ (Figure 3, page 79), the probability of an error of the second kind is about 0.42. On the other hand, if $\alpha = 0.01$ (Figure 4, page 80), the probability of this is 0.65. This means that, if the true value of the mean excess is as large as 1 percent and if we use alternatively $\alpha = 0.05$ and $\alpha = 0.01$, then in the circumstances of the experiments the mere existence of the advantage of a new variety over the standard will be detected in only about 58 or 35 cases, respectively, out of a hundred. From this you can see how the excess of caution with respect to errors of the first kind (0.01 in place of 0.05) leads to an increased chance of committing errors of the second kind (65 out of 100 in place of 42 out of 100).

Returning to the main subject of the conference, we notice that the graphs describing the dependence of the probability of errors of the second kind on the value of ρ and n are relevant from the point of view of the problems in plant breeding which we are considering. In practice, after a few years of existence, any seed breeding establishment must be aware of the size of the standard error per plot, say σ_0, which is likely to hold in future experiments. It is impossible to predict the exact value of σ_0, but it is certainly possible to make rough estimates of its upper limit. Therefore the breeder who contemplates experiments with m replications is able to substitute some reasonable number for σ into the expression for $\rho = \Delta/\sigma$, taking

$$\sigma = \sigma_0 \sqrt{\frac{2}{m}}. \tag{6}$$

He may then use tables or graphs of probabilities of errors of the second kind to find out what approximately will be his chance of detecting the advantage of his varieties when $\Delta = \overline{V} - \overline{S}$ has any value in which he may be interested. If he finds that, given a certain value of m, this chance is too small, then he will consider increasing the number m of replications. The increase of m will decrease the value of σ, increase the value of ρ, and consequently decrease the probability of an error of the second kind, i.e. the probability of failing to detect a good variety. This procedure must be considered essential in any rational planning of experiments.

But in the case of the plant breeder a special difficulty arises. Suppose he finds that, with five replications and $\alpha = 0.05$, the probability of detecting a good variety for which $\bar{V}$ exceeds the standard $\bar{S}$ by 5 percent is fairly large, say 0.9. It will be seen that this result is not very helpful. In fact, it is difficult to say beforehand how frequently his steps I through IV (page 85) will yield him new varieties which exceed the standard in sugar content by as much as 5 percent. It is possible that such success in breeding is unthinkable and that usually Δ does not exceed, say, one-third of a percent.

If one looks at the above mentioned graphs, it is easy to find that in such a case the chance of the breeder detecting the goodness of any of his varieties will be very small. Thus, if he keeps arranging his experiments with only $m = 5$ replications, practically all of his efforts in breeding new varieties will be wasted.

It is seen that the solution of the breeder's problem requires knowledge, not only of the probabilities of errors of the second kind, but also of the distribution of Δ in the population of new varieties which the breeder is likely to obtain in the future. It is impossible to predict what will happen in the future but it is possible to make rough guesses by studying what has happened in similar circumstances in the past. We may try to estimate the distribution of Δ in past years and use these estimates to obtain an idea of what may happen in the future.

The problem may be stated as follows. In some particular year, M experiments, comparing a large number N of new varieties with the same standard, gave N estimates, $x_1, x_2, \cdots, x_N$, of sugar excesses corresponding to the N varieties and M estimates, $s_1, s_2, \cdots, s_M$, of standard errors corresponding to the M experiments. It is required to use these numbers to estimate the distribution, say $p(\Delta)$, of the true excesses $\Delta_1, \Delta_2, \cdots, \Delta_N$, of the new varieties.

A similar problem was considered previously by Eddington and the solution is quoted by Levy and Roth.[3] However, Mrs. Tang offers a new approach. Her method consists of the following.

Denote by μ_k and ν_k the kth moments about zero of x and Δ respectively, and by σ^2 the variance of the experimental error ϵ in the observations x. If the traditional assumption is made that ϵ is normally distributed, then, as Mrs. Tang has calculated,

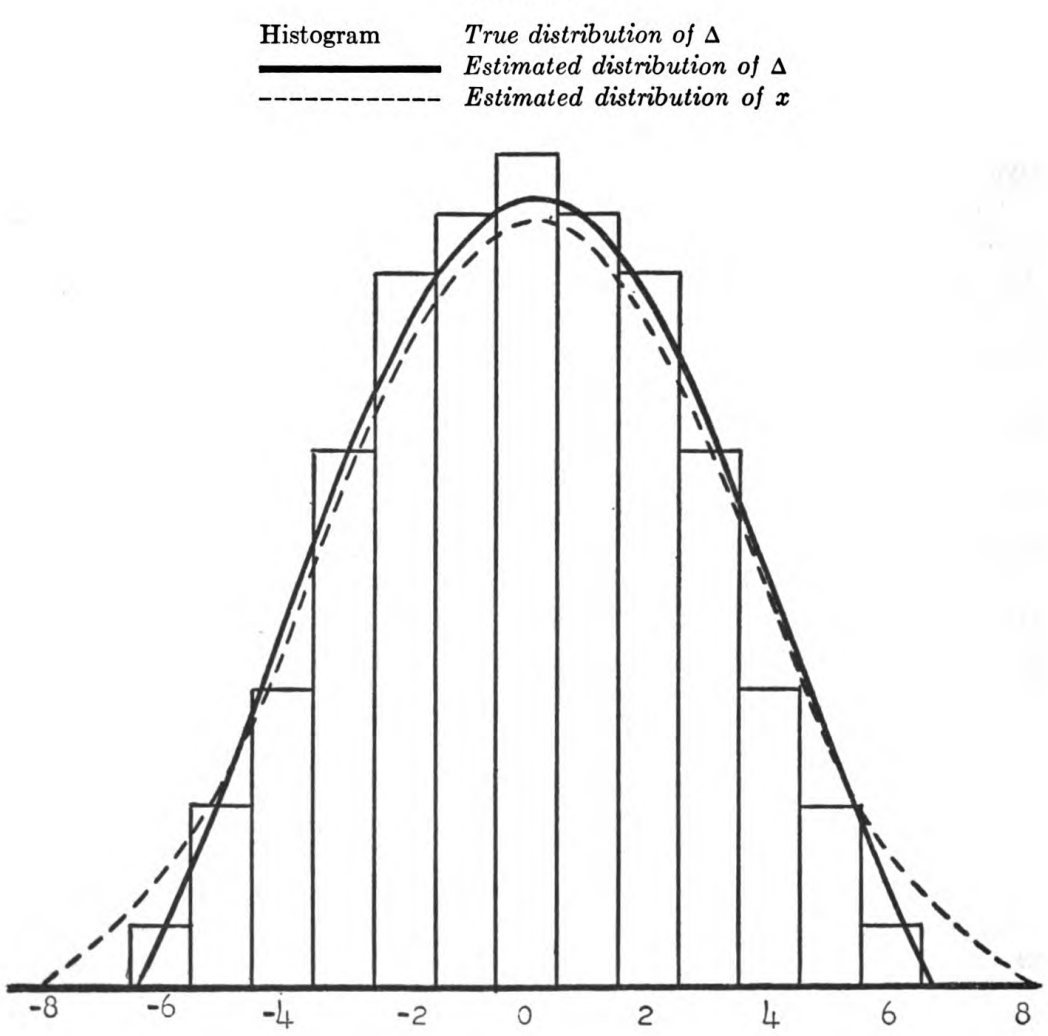
$$\left.\begin{aligned} \mu_1 &= \nu_1, \\ \mu_2 &= \nu_2 - \sigma^2, \\ \mu_3 &= \nu_3, \\ \mu_4 &= \nu_4 - 6\sigma^2\nu_2 + 3\sigma^4. \end{aligned}\right\} \tag{7}$$

[3] H. Levy and L. Roth, *Elements of Probability*. Oxford University Press, 1936, 200 pp.

Mrs. Tang also uses the assumption that σ has the same value in all M experiments. The use of this second assumption is partly justified by the fact that all the experiments are carried out by the same staff on the same large field with varieties which have many similar properties. The common value of σ can be estimated then with great accuracy since the estimate will be based on hundreds of degrees of freedom. This estimate, s', may be substituted in (7) for σ. Next, the observed values of x can be used to estimate the moments ν_1, ν_2, ν_3, ν_4. Together with s', they will yield the estimates of μ_1, μ_2, μ_3, μ_4. Finally, having obtained the μ's, Mrs. Tang uses them to fit a Pearson curve which is considered to be an estimate of $p(\Delta)$.

It is difficult to test the efficiency of this method theoretically. However, Mrs. Tang tried an empirical test. She started with an arbitrarily selected distribution represented by the histogram in Figure 2 (shown below). She

FIGURE 2



considered the histogram as the true distribution of N values of Δ in some possible two experiments. Next she used Mahalanobis' table [4] of normal deviates to produce values of x such as experiments with N new varieties, when her assumptions are satisfied, might have produced. In a similar way she obtained M values of the estimate of the error variance, each corresponding to one hypothetical experiment. After she had obtained these quasi empirical figures, she applied her method to estimate the distributions of Δ and x. Figure 2 shows the results. It is seen that the continuous curves do agree with the "true distribution" represented by the histogram.

QUESTION BY DR. SARLE: I am wondering what you used for a check.

ANSWER: I will explain it again. Let us assume that the true distribution of Δ is as follows:

Value of Δ	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Frequency	1	3	5	9	12	13	14	13	12	9	5	3	1

It is seen here that one of the Δ's is equal to -6, three others are equal to -5, etc. Write down the Δ in one column, thus:

$$\begin{aligned} \Delta_1 &= -6, \\ \Delta_2 &= -5, \\ \Delta_3 &= -5, \\ \Delta_4 &= -5, \\ \Delta_5 &= -4, \\ &\text{etc.} \end{aligned} \tag{8}$$

Next take from the table of Mahalanobis the corresponding number of values of ϵ; these values are so tabulated that they may be considered as values of a normal variate about zero with unit standard deviation. Suppose that you find

$$\begin{aligned} \epsilon_1 &= 0.03, \\ \epsilon_2 &= -1.16, \\ \epsilon_3 &= -0.25, \\ \epsilon_4 &= 0.53, \\ &\text{etc.} \end{aligned} \tag{9}$$

[4] P. C. Mahalanobis: "Tables of random samples from a normal population." *Sankhya*, Vol. 1 (1934), pp. 289–328.

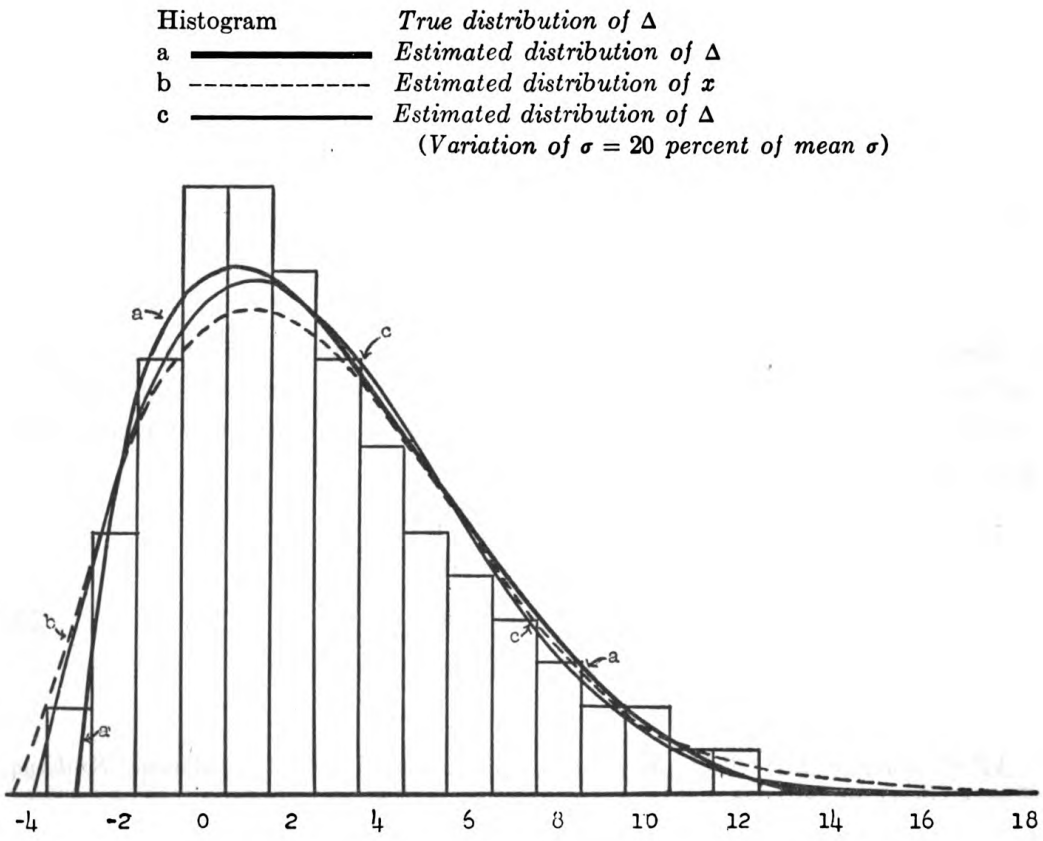
Now add these numbers to your Δ_i and you will obtain values which might be given by experiments if the true σ were unity and if the true Δ were distributed according to the above table. The results,

$$\left.\begin{aligned} x_1 &= -6 + 0.03 = -5.97, \\ x_2 &= -5 - 1.16 = -6.16, \\ x_3 &= -5 - 0.25 = -5.25, \\ x_4 &= -5 + 0.53 = -4.47, \\ &\text{etc.} \end{aligned}\right\} \qquad (10)$$

may now be used to estimate the distribution of x by the method of Mrs. Tang. Figure 2 represents the results.

You may have noticed that among the hypotheses of Mrs. Tang there is one that is doubtful. This is that the value of σ is the same in all experiments. Actually, in dealing with the results of real experiments it was found that this hypothesis may not be true. So Mrs. Tang checked, again empirically, that her method was still applicable with σ varying from one experiment to another within limits likely to occur in practice; Figure 3 shows the fit obtained with varying σ.

FIGURE 3



Having thus obtained an indication that her method does lead to reasonable results, Mrs. Tang applied it to the problem of estimating the distribution of true sugar excess over the standard in a number of new varieties tested in 1923 and 1924. The varieties were produced and tested by the breeders, K. Buszczynski and Sons, Ltd., of Warsaw, who kindly supplied the numerical data from their trials. Out of a considerable number of these trials, Mrs. Tang selected 40 carried out in 1923 and an equal number carried out in 1924. These were convenient as they had the same number of replications, namely 5. In each of the two sets, 120 new varieties were compared with the standard in a systematic arrangement like this:

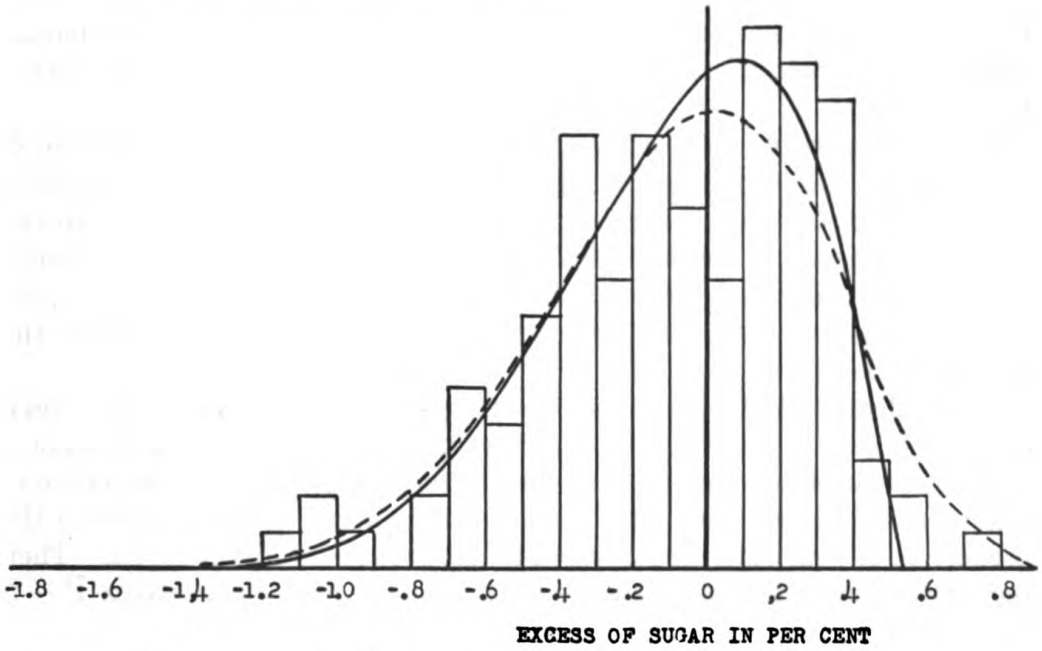
$$S\ V_1\ V_2\ V_3\ S\ V_1\ V_2\ V_3\ S\ V_1\ V_2\ V_3\ S\ V_1\ V_2\ V_3\ S\ V_1\ V_2\ V_3\ S \qquad (11)$$

To work out these experiments, i.e. to calculate the estimates x_i of the sugar excesses Δ_i, and the corresponding standard errors, Mrs. Tang applied the method of parabolic curves.[5] Next she estimated the distribution of Δ, the true sugar excess. Figure 4 gives the result referring to 1924. Here

FIGURE 4

Estimated distributions of sugar excess, 1924

Histogram — *Observed excesses of sugar content of 120 varieties over the standard*
---------- *Estimated excesses of sugar content of 120 varieties over the standard*
———————— *True excesses of sugar content*



[5] See conference on randomized and systematic experiments, pp. 67–84.

the histogram represents the observed distribution of x and the continuous curve the estimated distribution of Δ.

Similar curves calculated by the breeder may give him diversified and important information which I shall classify under two headings.

1. He may use such curves in analyzing his method of selecting parent plants, step II, page 85. If the breeder has recorded how he selected his plants a few years ago, he may usefully study what the distribution of Δ would have been if the selection had been made differently, say by breeding only half of the families that were actually taken. This would have allowed him to make a stricter selection of parent plants, taking only the very sweetest. If he ignores the new varieties that have been bred from the parent plants assumed to have been discarded in such cases and estimates the distribution of Δ for the remaining varieties, the breeder will be able to see whether or not the taking of many parent plants and the breeding of many new varieties does, in fact, represent a marked advantage.

2. When the breeder has the estimated distributions of Δ corresponding both to his actual experiments and also to the stricter method of selection at step II, he will be able to use the probabilities of errors of the second kind to see what the final results of his efforts, including step V, would have been. Let us illustrate this for the estimated distribution of Δ given in Figure 4.

The breeder is naturally interested in the varieties, conventionally called "good" varieties, for which $\Delta > 0$. Their proportion is represented by the area of the curve to the right of the origin. The breeder will be interested to know what proportion of "good" varieties is likely to be detected as "good" by his field trials when they are arranged according to this or that plan.

Take any positive value of Δ within the range of the curve in Figure 4, calculate the corresponding value of $\rho = \Delta/\sigma$ and determine the probability of an error of the second kind, corresponding to the value of ρ and to the number of degrees of freedom considered for the trials. Subtract this probability from unity and you will obtain the approximate value of the proportion $P(\Delta)$ of good varieties that will be detected as "good" by the proposed trial.

Now calculate $P(\Delta')$ for a number of successive values Δ' of Δ. Next, for these values Δ', take the estimated ordinate $p(\Delta')$ of the distribution of Δ in the population of your new varieties (as for example, the full line curve of Figure 4). This ordinate multiplied by $\delta\Delta$ is approximately equal to the proportion of your varieties for which Δ falls between Δ' and $\Delta' + \delta\Delta$. Then the proportion of the new varieties that (a) have their sugar excess $\overline{V} - \overline{S}$ between Δ' and $\Delta' + \delta\Delta$, and (b) will be detected as good varieties by the field trials planned will be obtained by the multiplication

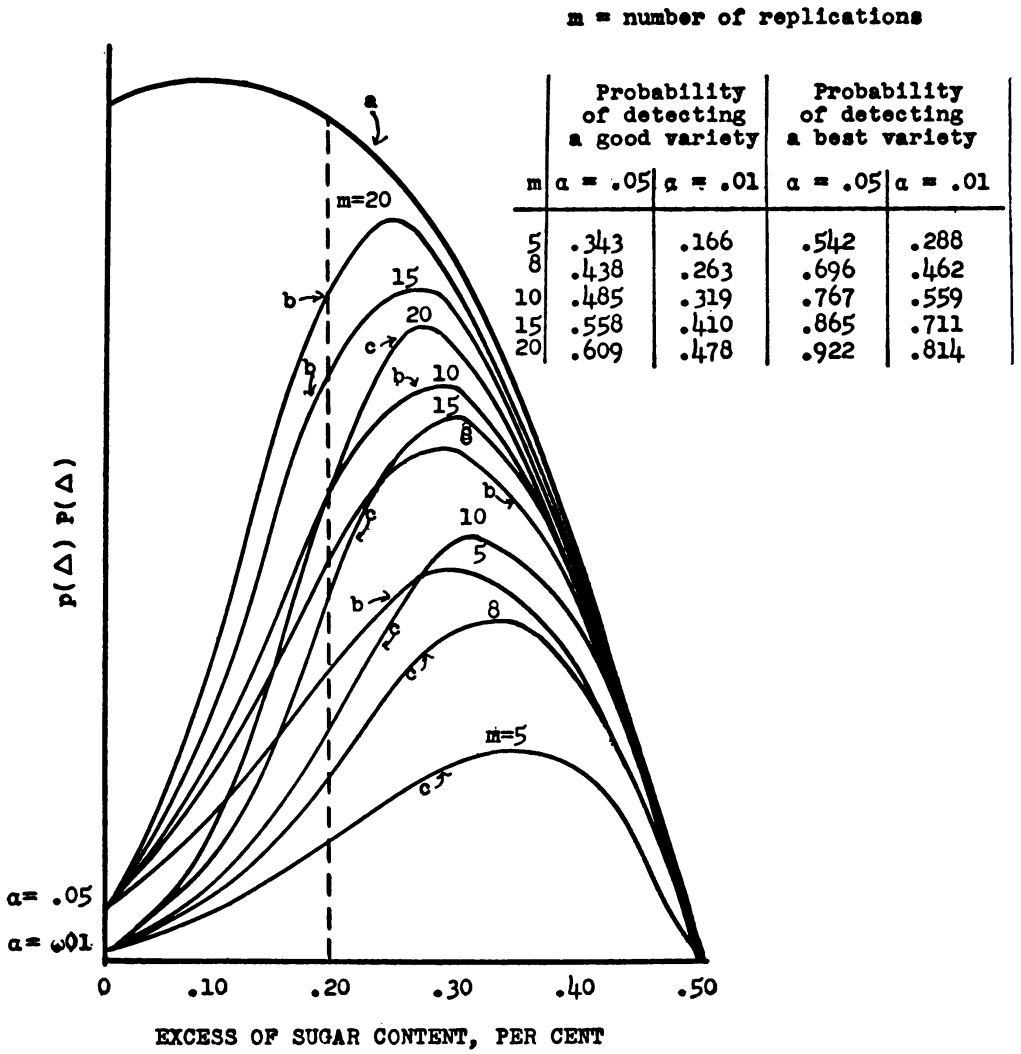
$$p(\Delta')P(\Delta')\delta\Delta.$$

Figure 5 was constructed in this way for the good varieties of Figure 4, i.e., it was made up from that part of the estimated distribution of Δ in Figure 4 lying to the right of the origin. The uppermost curve a of Figure 5 is simply the full line curve lying to the right of the origin in Figure 4. The dimensions are reduced so that the area under this part of the curve

FIGURE 5

Distributions of true sugar excess

a *In population of varieties tested*
b *In population of varieties found significant at* $\alpha = .05$
c *In population of varieties found significant at* $\alpha = .01$



is equal to unity: we are interested only in "good" varieties and in the proportion likely to be detected as such.

The lower curves labeled b and c represent plotted values of the products $p(\Delta)P(\Delta)$, where $P(\Delta)$ corresponds to $\alpha = 0.05$ or 0.01 and to different arrangements of the proposed experiments. It was assumed that all the experiments had been arranged in randomized blocks and differed only in the number of replications m, marked on each curve. The curves b corresponding to $\alpha = 0.05$ end at the point 0.05 on the axis of ordinates. The other curves c corresponding to $\alpha = 0.01$ have this ordinate equal to 0.01.

The area under each curve represents the proportion of "good" varieties that will be recognized as such for a given α and a given number m of replications. In addition the curves give the distribution of Δ for the "good" varieties that will be detected. You will see that if the stricter level of significance $\alpha = 0.01$ is applied and if the number m of replications is as small as 5, then the proportion of good varieties that will be detected is very small. You will find its value, 16.6 percent, on the small table attached to Figure 5, page 99. This number 16.6 percent, is the area under the curve for $\alpha = 0.01$ and $m = 5$, divided by the area under the curve marked a. On the other hand, if $\alpha = 0.05$, then the same proportion rises to 34.3 percent. If the number of replications is doubled, then the corresponding figures will be 31.9 and 48.5 percent, respectively.

Apart from the proportion of good varieties likely to be detected, the breeder may be interested in the proportion of those for which the value of Δ is not merely positive but exceeds some arbitrary limit, say 0.2 percent of sugar. Such varieties may conventionally be termed the "best." There is no difficulty in calculating the proportions of the "best" varieties whose superiority over the standard would be detected by the trials. We have only to use the areas of all the curves to the right of the line $\Delta = 0.2$. The corresponding figures are given in the two "best" columns of the table attached to Figure 5. For instance, in the table under "Probability of detecting a *best* variety," at $\alpha = 0.05$ and $m = 8$, we see 0.696. This means that the area to the right of 0.2 percent under the curve for $\alpha = 0.05$ and $m = 8$ is 0.696 of the area to the right of 0.2 percent under the curve marked a.

Figure 5, the table and the method of construction represent the main result of the work of Mrs. Y. Tang. The breeder who now starts 500 new varieties each year and replicates them only 5 times in his trials may use her results to construct curves similar to those in Figures 4 and 5, and may thus compare the probable results of his work for the cases in which he started, not with 500 families, but perhaps with 400, 300, 200 and a corresponding increase in the number of replications. Having these results before his eyes he will be able to take into account various economic

factors and choose the most economical relation between the number of replications and that of the new families started.

I might conclude here, but it seems advisable to warn the reader that the actual process of seed breeding is a little more complex than that presented above. In fact, it is extremely difficult to include in formulas an exact process of any more or less complicated practical work. This is also the position in the present case. In order to give an idea of what I have in mind, I may remind you of one thing that I have already mentioned—new varieties are tested for more than one vegetative period and in more than one spot. It follows that the method built by Mrs. Tang refers to a simplified case. But also it is obvious that she has contributed to our technique by showing how to calculate the probable results of only one series of field trials when no such method existed before. And even though this is not all that is needed, it is a great deal because the most difficult part of any problem consists in noticing that there *is* a problem and in advancing some sort of solution. There are usually a lot of people able to introduce the necessary corrections and extensions.

QUESTION BY DR. SARLE, POINTING TO FIGURE 5: What basis do we have for figuring the possibility of including some "false good" varieties in this area? Will all poor ones be eliminated by this process, or is there a chance of getting some of the poor ones?

ANSWER: Figure 5 refers only to those varieties that are really "good." The control of "false good" varieties is kept by choosing a proper level of significance. If you fix $\alpha = 0.05$, then the chance that the best out of the "false good" varieties (those with $\Delta = 0$) will be passed as good is 0.05. On the other hand, the areas under sections of the curves in Figure 5 give the proportions of the varieties that are really "good."

QUESTION BY DR. SARLE: Your method automatically does that?

ANSWER: In principle, yes: but we must remember that the method gives only an estimate which is always liable to error.

QUESTION BY DR. SARLE: How does it know which one to pick out?

ANSWER: It doesn't. It would be a great thing if it did. All that it can do is to estimate proportions. If you toss a fair penny, you can never tell exactly when it will fall heads. On the other hand, you can safely say that in the long run the proportion of heads will be about one-half. Similarly, no statistical method is able to indicate which of the varieties with positive x is really "good" and which is "false." On the other hand it is possible to estimate the proportion of those that are really "good" and also the proportion of their number which will be detected as "good."

QUESTION BY DR. SALMON: This means that with five replications you actually identify only a relatively small percentage of the total number of good varieties.

ANSWER: Yes, a very small percentage. But we must remember that the accuracy of experiments varies a great deal from year to year, owing to weather conditions. As a matter of fact, in the year 1923 which was also studied by Mrs. Tang, the proportion was found to be much greater than that indicated here.

CHAPTER III

Some Statistical Problems in Social and Economic Research

(This chapter is dedicated to the memory of Dr. Kazimierz Karniłowicz, the late Director of the Institute for Social Problems, Warsaw, Poland.)

Part 1. Sampling Human Populations. General Theory.

The present section is based on a conference held in the Auditorium of the United States Department of Agriculture, April 8, 1937, Dr. Frank M. Weida presiding. At this conference, I summarized my general ideas on sampling human populations and gave some theoretical results which I had obtained in connection with a sampling survey of Polish labor conducted by the Institute for Social Problems in Warsaw. The methodology developed was originally published in Polish.[1] Later on, the main theoretical results were incorporated into a critical survey of sampling methods, published in English.[2] The numerical results obtained in the sampling survey by the Institute for Social Problems were published by Jan Piekałkiewicz.[3]

The subject of the conference of April 8, 1937, was selected as a result of numerous letters received from prospective members of the audience. All these letters visualized the following general situation: a certain amount of money is available for a survey and the problem is to determine the sampling procedure which will make the best use of these funds. Such differences as were present in the several particular problems described in the correspondence referred to specific circumstances of sampling and to special characteristics of the population studied. The purpose of the conference was to develop some general ideas from which answers to a number of particular questions could be derived.

One of my correspondents had in mind a population of 300 cities. In order to study this population he intended to select a sample of 25 cities

[1] J. Neyman: *An Outline of the Theory and Practice of the Representative Method Applied in Social Research.* Institute for Social Problems, Warsaw, 1933, 123 pp. (Polish).

[2] J. Neyman: "On the two different aspects of the representative method." *Jr. Roy. Stat. Soc.,* Vol. 97 (1934), pp. 558–625.

[3] Jan Piekałkiewicz: *Report on the Study of the Structure of Polish Labor.* Institute for Social Problems, Warsaw, 1934, 238 pp.

to be covered by a 100 percent enumeration. His question was how best to select 25 cities from the 300 so that he could draw conclusions regarding the inhabitants of all the 300 cities under investigation.

I am not going to answer this question. Instead I am going to advise as strongly as I can that the proposed method of sampling be dropped altogether. This method is most dangerous and is practically certain to lead to deplorable results. Of course, I do not mean by this that a successful inquiry by means of sampling is impossible. On the contrary, it is my opinion that the sampling method is useful and can provide very accurate results. What I emphatically protest against is the selection of any 25 cities for a complete census (100 percent enumeration) with the consequent total omission of the remaining 275 cities.

Broadly speaking, there are two essentially different methods of sampling used in social work. One is called the method of *purposive selection*, the other that of *random sampling*. This subdivision is a little artificial, but owing to the fact that it was used in a special report [4] on the method, presented to the International Statistical Institute, it is generally accepted.

The method which consists in selecting 25 out of 300 cities and in limiting the investigation to these 25 cities falls under the heading of "purposive selection." The mere question of how one should best select these cities suggests that the selection was not meant to be random, at least not entirely random. Usually it is suggested that the sample of cities should be selected so that the averages of certain characters, called controls, calculated for the sample and for the universe should be in as close agreement as possible. It is this circumstance which justifies the term "purposive selection." But it is not the limitation of the randomness of sampling which makes the method dangerous. In fact, if the question concerned only random sampling, I could easily answer it by saying that the best way of selecting the 25 cities is to draw them at random.

The trouble with the method lies in the fact that if we try to select things (cities, districts, etc.) "purposely," then both the total number of units from which selection is to be made, and even more inevitably, the number selected must necessarily be small, and therefore the units themselves must be rather large. In the present case we have 300 units out of which only 25 are to be selected. Each unit of selection is a city inhabited probably by tens of thousands of people, possibly more, and the differences between the units may be enormous. This is a rough description of the method called "purposive selection."

The nomenclature "purposive selection" and "random sampling" is not very felicitous, as I have already indicated. It does not describe the

[4] L. A. Bowley: "Measurement of the precision attained in sampling." *Bull. Inst. Int. Stat.*, Vol. 22 (1925), 1er livre, pp. 1–62, supplement.

essential difference between the two methods when they are applied in practice. The first method, that of "purposive selection," consists in dividing the whole population into a comparatively few (say 300) large groups (e.g. cities) or units, of which some 20 or 30 are selected "purposely." The essential feature of the other method is that the same population is divided into a much larger number (say 100,000 or more) of small groups (e.g. families, inhabitants of single houses, blocks, etc.) from which a sample of around 1000 or more are selected, either entirely at random or at random with some restrictions.

The first method is hopeless, the other extremely useful. If anyone would like to see theoretical reasons for this opinion, he will find them in my paper published in the *Journal of the Royal Statistical Society,* already quoted. In this conference I will give an intuitive illustration of the ideas expressed there. Suppose we have a hundred dollars that we decide to use for gambling in a fair game. If we divide the whole sum into, say, five parts of $20 each and bet only five times, it is impossible to make a reliable prediction of what the result may be. We may lose all our money, or equally easily, we may double it. On the other hand, if we make a hundred bets at $1 each, then we can make some predictions with fair hope of success. The result of the game still remains uncertain, but it would be rather surprising if the sum won or lost exceeded $20. The accuracy of the prediction would be still greater if, instead of making a hundred bets at $1, we would make a thousand bets of a dime.

These are perfectly intuitive propositions and you will notice that they have a definite bearing on the problem of sampling human populations. The advice against selecting 25 cities out of a total of 300 is not based on theoretical considerations alone: some practical experience is available to show what the result of an inquiry might be if this method is applied.

In 1926 or 1927 two Italian statisticians, Gini and Galvani,[5] had to solve a problem of a kind that is exactly similar to the one contemplated here. They had to deal with the data of a general census. The data were worked out, a new census was approaching, and the room had to be cleared for the new data. The old data were to be destroyed, but the statistical office wanted to keep a representative sample so as to have material for future studies, as yet unanticipated. Gini and Galvani were responsible for the method of obtaining a sample which would represent the situation in the whole of Italy. What they did is a good example of how not to sample human populations.

The two authors carefully considered the problem, took into account

[5] Corrado Gini and Luigi Galvani: "Di una applicazione del metoda rappresentativo all'ultimo censimento italiano della popolazione." *Annali di Statistica,* Serie vi, Vol. 4 (1929), pp. 1–107.

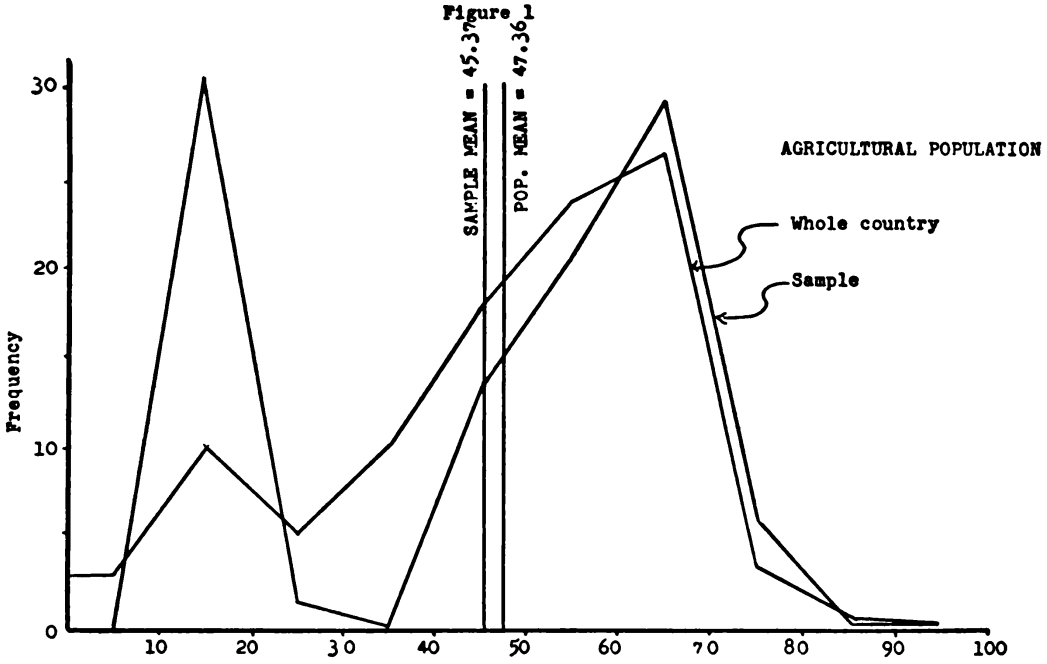
the Report of the International Statistical Institute and decided to apply the method of purposive selection. The whole of Italy was divided into 214 administrative districts called *circondarî* and out of these 29 *circondarî* were selected to form the sample. Some of the *circondarî* are large districts with more than a million inhabitants. It is interesting to note that the ratio of the Italian sample to the universe sampled, 29:214, is substantially larger than the sampling ratio contemplated by my correspondent, 25:300.

Various averages for each *circondario* had been calculated previously. Gini and Galvani selected 12 characters of the *circondarî* to serve as controls and subdivided these into essential and secondary controls. They tried to select the 29 *circondarî* so that the means of the essential controls calculated from the sample would be practically identical with those for the whole population. They also tried to reach a reasonable agreement between the population and the sample means of the secondary controls. If you will look at the figures, you will find that the agreement of the mean of each control in the sample with the mean of the same control in the population is very good.

From the paper by Gini and Galvani, it is uncertain whether or not the old Italian census data were destroyed and the sample was left for future reference. However, the two authors decided to check the goodness of the sample by comparing its various characteristics with those known for the whole population of Italy. The results of this comparison are described by Gini and Galvani and should be kept in mind as an argument against the use of the purposive selection method. Gini and Galvani found that the distributions of various characteristics of the individuals, the correlations, and, in fact, all statistics other than the average values of the controls showed a violent contrast between the sample and the whole population. Figure 1 reproduces a diagram taken from page 95 of the paper by Gini and Galvani, which illustrates the situation. You will see that the distribution observed in the sample bears little resemblance to that of the whole population.

Having discovered that their sample of 29 *circondarî* is not at all representative of the whole population, the Italian statisticians expressed the opinion that, generally, it is impossible to obtain a sample that reproduces the population sampled and all its properties. Strictly speaking, they are correct. In 1926 there was in Italy but one Marchese Marconi, the great inventor in the field of wireless telegraphy. Whatever the method of sampling, the proportion of Marconis in the sample can not be equal to that in the population. But we do not take samples to establish such proportions; and both theory and experience indicate that, whenever we have in mind a truly statistical problem of estimating means of any size, of

Figure 1



PROPORTION OF POPULATION ENGAGED IN AGRICULTURE

(Taken from page 95 of Gini and Galvani's article in the *Annali di Statistica,* Serie vi, Vol. 4, 1929)

regressions, etc., a *properly drawn* sample is, for all practical purposes, sufficient.

Now let us consider what is to be done to get a reliable sample. Here we must rely on the theory of probability and work with great numbers. "Great numbers" does not mean great numbers of people included in the sample, but great numbers of random selections to form a sample, or great numbers of units that are drawn separately. The sample of 25 cities or the sample of 29 *circondarî* contain a great number of people, but from the point of view of sampling theory *they are both small samples* because they are composed of 25 or 29 units, respectively. For a sample to be reliable the number of units must be large.

Thus, instead of dividing your population into 300 parts, each representing a particular city, you need to carry the subdivision much farther. Probably it would be best to divide the whole population of 300 cities into small groups inhabiting single houses or blocks. All these groups, which I shall call units of sampling, or simply units, must be listed, and the necessity of listing usually imposes a limit to the tendency of having the units very small.

When the population sampled is represented by the mass of records

obtained by a general census, then the smallest unit you can choose conveniently is the smallest division for which there is a separate folder in your data, or for which you have a separate punch card. This division may represent a block, a household, etc., the smaller the better. Ordinarily, since such divisions are small, there is a great number of them in the population studied. Therefore no great difficulty occurs in sampling from existing records. The situation is much more difficult if you are to sample people, not records already collected. In this case you have to send enumerators into the field and give them addresses at which to call. In order to insure randomness of the sample, the addresses must be selected truly at random and this requires a previously compiled list of all addresses forming the population. Frequently such a list is unavailable and this causes considerable difficulty. However, this difficulty may be overcome in part by using a map that divides the area under investigation into a large number of small sections and by considering each section as a unit of sampling. Whatever the selected unit, it is a relatively simple matter to produce a random sample of any preassigned size once you have a complete list of the units forming the population.

Several questions addressed to me were concerned with what proportion the size of the sample should bear to the size of the population. This proportion does affect the precision but in a much milder way than the number of units selected to form the sample. Thus, a sample of 10,000 units (blocks, inhabitants of separate houses, etc.) will be very accurate almost irrespective of whether it forms 10 percent of the population studied or one percent or one-tenth of a percent.

The process of random sampling may be of various forms which are not equivalent from the point of view of the accuracy of the results. The first attempt at a serious study of the relation between the method of sampling and the accuracy of results was made by Bowley and is described in his report to the International Statistical Institute already mentioned. The main results of his study are as follows.

Random sampling is called unrestricted if at each drawing each of the elements forming the population studied has the same chance of being drawn. To illustrate this idea I shall point out that, if the population is formed by the inhabitants of 300 cities and if the unit of sampling is represented by a block, then unrestricted sampling combined with bad luck can produce a sample composed of blocks from just one city with the complete omission of other cities. However this is extremely unlikely.

More accurate results could be obtained by what Bowley calls stratified sampling and what I call *stratified proportional sampling*. This consists in a twofold subdivision of the population studied. First, we divide it into a convenient number of larger parts, called strata. For example, a

stratum may be a city or a large section of a city. Next, each stratum is divided into units of sampling. If you have decided to work with a sample of one-twelfth of the population, then from each stratum separately you select at random one-twelfth of its units. This makes it impossible for the sample to be devoid of units representative of larger sections of the population studied.

When we divide the population into strata, we should remember that the more homogeneous the single strata, the better will be the effect of stratification. In practically every city certain sections are easy to distinguish as those inhabited by the well-to-do, those of poorer people, shopping districts and industrial areas. In order to achieve better accuracy, each of these sections should be treated as a separate stratum.

However, homogeneity of a stratum does not necessarily mean equality or similarity of all people inhabiting this stratum. In fact, homogeneity of a stratum or of a population means a comparative similarity of the *units of sampling,* rather than of the individuals forming the units. If the population of a town is composed of representatives of ten different races, each in the same proportion, then we would say, probably, that this population is very heterogeneous from the racial point of view. However, from the point of view of sampling, this population would be ideally homogeneous if it happened that the racial composition of each of its sampling units is exactly the same as that of the whole population. Thus one sees that the internal heterogeneity of sampling units goes with an external homogeneity of these units within the population. This is a general rule.

From this it follows that the choice of sampling units of a fixed size is not indifferent from the point of view of the accuracy of an investigation by sample. Dr. Frederick F. Stephan tells me that an investigation has shown the existence of a greater similarity between the inhabitants of two sides of one street than between those of opposite sides of the same block. Hence, if one contemplates dividing the population alternatively into units of sampling composed of the inhabitants of the two sides of sections of single streets or of the two sides of single blocks, the latter method would give more homogeneous units and therefore greater accuracy of sampling.

Frequently, the gain in accuracy resulting from stratification is considerable, but it is possible to go further than Bowley advised. A cursory glance at the situation suggests that the rule of selecting randomly *the same* proportion of units out of each stratum may not be the best procedure. You can not expect that all the strata will be equally homogeneous internally. To make the situation clear, suppose that one of the strata, A, is ideally homogeneous, while another, B, is fairly heterogeneous. Then, in order to know all about the stratum A, it is sufficient to take a sample of only one unit. On the other hand, an accurate estimate of the properties of B would

require a sample of considerable size. If we decide to sample both A and B in proportion to their sizes (= the number of elements of sampling they contain), then we shall "oversample" A and "undersample" B. This intuitive reasoning is fully supported by the theory I developed in my article of 1934, already quoted.

After this somewhat general discussion, let me enter into a few details. Consider a population, say Π, divided into a certain number s of strata, say $\pi_1, \pi_2, \cdots, \pi_i, \cdots, \pi_s$. Further, assume that the ith stratum contains M_i units of sampling, numbered from 1 to M_i, and let

$$u_{i1}, u_{i2}, \cdots, u_{ij}, \cdots, u_iM_i \tag{1}$$

be the values of a certain numerical characteristic U of these units. Our problem is to estimate the grand average $U..$ referring to the whole population Π. In other words, if M_0 stands for the total number of units of sampling in the whole population,

$$M_0 = \sum_{i=1}^{s} M_i, \tag{2}$$

then

$$U.. = \frac{1}{M_0} \sum_{i=1}^{s} \sum_{j=1}^{M_i} u_{ij} = \frac{1}{M_0} \sum_{i=1}^{s} M_i u_{i\cdot}, \tag{3}$$

where $u_{i\cdot}$ denotes the average value of the characteristic U relating to the ith stratum.

While the general situation is being considered, it is convenient to have in mind one or two specific examples. Thus, the purpose of a certain sampling survey may be to establish the total number of unemployed in a given area A of the country. In this case the area A may be divided into s sections representing strata. Each stratum may be sub-divided into a number of convenient sampling units, e.g., blocks or merely squares on the map. The symbol u_{ij} will denote the number of unemployed inhabiting the jth block of the ith stratum. Once we have estimated the grand average $U..$ of the number of unemployed per block, it is a simple matter, if we know the number of blocks, to estimate the total number of unemployed in the whole area A.

Alternatively, the purpose of the sampling survey may be to estimate the average expenditure on housing (or any other item) per family of unemployed inhabiting the area A. This problem is a little more complicated because it splits into two: (1) to estimate the total expenditure on housing of all the unemployed and (2) to estimate the total number of families of unemployed. Thus, actually, we have a combination of two related problems but I intend to discuss in detail only a single problem. In this case it would be the problem of estimating the average expenditure

on housing of unemployed families per one unit of sampling. Here u_{ij} would mean the combined expenditure on housing of families of unemployed that inhabit the jth block of the ith stratum. If the total number of unemployed is known or if it is estimated by the same inquiry, then the knowledge of the average per block of expenditures on housing will provide the requisite average per family.

Returning to the general theoretical case, denote by m_i the number of units which we intend to select from the ith stratum for $i = 1, 2, \cdots, s$, to form a sample on which to base an estimate of $U\ldots$ Denote by X_{ij} the value of the characteristic U to be observed in the jth unit of sampling selected from the ith stratum. Before the sample is actually drawn, the exact values of X_{i1}, $X_{i2}, \cdots, X_{im_i}$ are unknown and any one of the numbers (1) may appear as the value of X_{i1}, any one of these numbers may appear as the value of X_{i2}, etc. In fact, the symbols $X_{i1}, X_{i2}, \cdots, X_{im_i}$ may represent any one of the many combinations of m_i out of the M_i numbers (1). Before the sample is drawn, the X_{ij}'s are random variables.

Let

$$X_{i\cdot} = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{ij}. \tag{4}$$

Then the best linear unbiased estimate [6] of the grand average $U..$ is

$$X.. = \frac{1}{M_0} \sum_{i=1}^{s} M_i X_{i\cdot}. \tag{5}$$

It is customary to measure the precision of this estimate by the value of its variance, say σ^2. The smaller the variance is, the better the precision. The theoretical problem before us is to determine the best way of using the funds available for the survey in order to minimize the variance σ^2. The first solution of this problem is contained in my article already quoted. However, this solution applies to the special case where the cost of sampling or, more precisely, the average cost per unit of sampling is the same in all the s strata. This condition is frequently satisfied. In some cases, however, when certain of the strata are urban and others are rural, the average cost of sampling a unit may vary considerably from stratum to stratum. Since the method of determining the stratification which will be optimum in the use of sampling funds is exactly the same whether the average cost is constant or not, we shall consider the more general case.

Assume then that the total expenditure on the survey is fixed and is

[6] The term "best linear unbiased estimate" is very intuitive and familiar to many statisticians. Details of the definition and some theory may be found in the excellent little book by F. N. David: *Probability Theory for Statistical Methods.* Cambridge University Press, 1949, 230 pp.

equal to C dollars. Further let c_i represent the average cost per unit of sampling in the ith stratum. Then the numbers m_i, $i = 1, 2, \cdots, s$, of units to be selected from the whole population Π must satisfy the condition

$$m_1c_1 + m_2c_2 + \cdots + m_ic_i + \cdots + m_sc_s = C. \tag{6}$$

Our problem is to determine the numbers m_i so as to satisfy (6) and so as to minimize the variance σ^2. The value of σ^2 corresponding to any fixed system of the m_i is obtained from the formula

$$\sigma^2 = \frac{1}{M_0^2} \sum_{i=1}^{s} M_i \frac{M_i - m_i}{m_i} \sigma_i^2, \tag{7}$$

where

$$\sigma_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (u_{ij} - u_{i\cdot})^2 \tag{8}$$

represents the internal variability within the ith stratum. Formula (7) must be familiar [7] to many of you so I shall not bother you with its proof. However formula (7) is not convenient for our purposes since it does not immediately bring out the effect of choosing this or that system of values of the m_i, which determines the stratification of the sample. For this reason we shall rewrite formula (7) in an alternative form. This is obtained by using the familiar identity applicable to any numbers $\alpha_1, \alpha_2, \cdots, \alpha_s$ without any restriction and to any "weights" $w_1, w_2, \cdots, w_s$, provided the sum of the weights is different from zero. The identity in question is

$$\sum_{i=1}^{s} w_i\alpha_i^2 = \alpha_\cdot^2 \sum_{i=1}^{s} w_i + \sum_{i=1}^{s} w_i(\alpha_i - \alpha_\cdot)^2 \tag{9}$$

where

$$\alpha_\cdot = \frac{\sum_{i=1}^{s} w_i\alpha_i}{\sum_{i=1}^{s} w_i} \tag{10}$$

is the weighted mean of the α's.

Returning to formula (7), we split the right hand side into two parts of which only the first depends on the m_i,

$$\sigma^2 = \frac{1}{M_0^2} \left\{ \sum_{i=1}^{s} \frac{M_i^2\sigma_i^2}{m_i} - \sum_{i=1}^{s} M_i\sigma_i^2 \right\}. \tag{11}$$

Now consider the first sum within the curved brackets. Multiply and divide the ith term of this sum by the product m_ic_i and apply formula (9),

[7] This formula is deduced in detail in the book by F. N. David, already quoted.

letting $m_i c_i$ play the role of the weights w_i and the quotients

$$\frac{M_i \sigma_i}{m_i \sqrt{c_i}} \tag{12}$$

the role of the arbitrary numbers α_i. Remembering condition (6) we have

$$\sum_{i=1}^{s} \frac{M_i^2 \sigma_i^2}{m_i} = \sum_{i=1}^{s} m_i c_i \left(\frac{M_i \sigma_i}{m_i \sqrt{c_i}} \right)^2$$

$$= CA^2 + \sum_{i=1}^{s} m_i c_i \left(\frac{M_i \sigma_i}{m_i \sqrt{c_i}} - A \right)^2, \tag{13}$$

where

$$A = \frac{1}{\sum_{j=1}^{s} m_j c_j} \sum_{i=1}^{s} m_i c_i \left(\frac{M_i \sigma_i}{m_i \sqrt{c_i}} \right)$$

$$= \frac{1}{C} \sum_{i=1}^{s} M_i \sigma_i \sqrt{c_i} \tag{14}$$

is the weighted mean of the quotients

$$\frac{M_i \sigma_i}{m_i \sqrt{c_i}}$$

and appears to be a fixed number, independent of the m_i. Substituting (13) into (11), we obtain the desired formula for the variance σ^2,

$$\sigma^2 = \frac{1}{M_0^2} \left\{ CA^2 + \sum_{i=1}^{s} m_i c_i \left(\frac{M_i \sigma_i}{m_i \sqrt{c_i}} - A \right)^2 - \sum_{i=1}^{s} M_i \sigma_i^2 \right\}. \tag{15}$$

In virtue of (14) only the second term in the curved brackets of (15) depends on the numbers m_i which determine the stratification of the sample. This term is a weighted sum of squares of differences between the quotients

$$\frac{M_i \sigma_i}{m_i \sqrt{c_i}}$$

and their weighted mean A. It follows that, in order to minimize σ^2, it is both necessary and sufficient to ascribe values, say m_i^*, to the m_i such that for each i

$$\frac{M_i \sigma_i}{m_i^* \sqrt{c_i}} = \frac{1}{C} \sum_{j=1}^{s} M_j \sigma_j \sqrt{c_j} = A. \tag{16}$$

This implies that the optimum stratification of the sample is determined by

$$m_i^* = \frac{C}{\sum_{j=1}^{s} M_j\sigma_j\sqrt{c_j}} \frac{M_i\sigma_i}{\sqrt{c_i}}, \quad i = 1, 2, \cdots, s. \tag{17}$$

Note added in proof: A result equivalent to formula (17) is contained in the book, *Some Theory of Sampling,* by William Edwards Deming (John Wiley and Sons, New York, 1950), which appeared between the time the above lines were written and the reading of the proofs. However, Deming's method of obtaining the result is different from the one used here.

It is seen that the numbers m_i^* so determined automatically satisfy condition (6) which states that the total cost of sampling must be C.

In practice it will be impossible to satisfy formula (17) exactly because the numbers m_i^* must be integers while ordinarily the right hand side of (17) is an irrational number. However, this difficulty is trivial and the optimum stratification of the sample may be taken as the system of s integer numbers closest to the values of the right hand side of (17) for $i = 1, 2, \cdots, s$, or just exceeding them. With this stratification, we may ignore the middle term in (15) and write, say,

$$\sigma^2_{\text{opt}} = \frac{1}{M_0^2}\left\{\frac{1}{C}\left(\sum_{i=1}^{s} M_i\sigma_i\sqrt{c_i}\right)^2 - \sum_{i=1}^{s} M_i\sigma_i^2\right\} \tag{18}$$

(with a very good approximation) as the minimum variance of the estimate of $U..$ attainable with the optimum stratification of the sample.

Returning to formula (17), you will notice that, roughly speaking, in order to attain the greatest benefit from a given stratification of the population, you should sample more heavily the strata which are more variable and also the strata in which sampling is less expensive.

In order to see the effect of sampling proportionately to the sizes M_i of the strata, put $m_i = kM_i$ where k stands for the factor of proportionality. This factor is determined from the condition that the total cost of sampling must be C,

$$\sum_{j=1}^{s} m_jc_j = k\sum_{j=1}^{s} M_jc_j = C \tag{19}$$

so that

$$m_i = kM_i = M_i\frac{C}{\sum_{j=1}^{s} M_jc_j}. \tag{20}$$

Denoting by σ^2_{prop} the variance of the estimate $X..$ corresponding to the proportional system of sampling and using (15) and (18), we have

$$\sigma^2_{\text{prop}} = \sigma^2_{\text{opt}} + \frac{1}{M_0^2}\frac{\sum_{j=1}^{s} M_j c_j}{C}\sum_{i=1}^{s} M_i c_i\left(\frac{\sigma_i}{\sqrt{c_i}} - B\right)^2, \tag{21}$$

where B stands for the weighted mean of the quotients $\sigma_i/\sqrt{c_i}$,

$$B = \frac{1}{\sum_{j=1}^{s} M_j c_j}\sum_{i=1}^{s} M_i c_i \frac{\sigma_i}{\sqrt{c_i}} = \frac{AC}{\sum_{j=1}^{s} M_j c_j}. \tag{22}$$

The importance of the disadvantage of proportional sampling when compared to the optimum depends, then, on the variability of the quotients $\sigma_i/\sqrt{c_i}$. If the population sampled contains the whole of a geographic area one may expect that σ_1 will be larger in the urban districts than in the rural. On the contrary, one may expect the values of the c_i, on account of the cost of travel, to be smaller in the cities than in the country. Thus, both factors considered are likely to contribute to the variation of the quotients $\sigma_i/\sqrt{c_i}$. As a result, it seems probable that, in order to attain the precision in the estimate which is best within the limits of funds available, the rural areas should be sampled less heavily than the cities.

Frequently one hears the assertion that, whatever way one stratifies a population, proportional sampling will give results which are always more precise than an unrestrictedly random sample of equal size. It is important to remember that this assertion is false. To show this, let us consider the simplest case where the cost of sampling per unit is exactly the same in all parts of the population.

If we ignore the stratification of the population Π and base the estimate of one grand mean $U..$ on an unrestrictedly random sample of m_0 units drawn from Π, then the variance of the estimate, say σ_u^2 will be represented by just the term similar to the general term in (7), namely,

$$\sigma_u^2 = \frac{M_0 - m_0}{m_0 M_0(M_0 - 1)}\sum_{i=1}^{s}\sum_{j=1}^{M_i}(u_{ij} - U..)^2. \tag{23}$$

By adding and subtracting $u_{i\cdot}$ within the parenthesis and then expanding the square, expression (23) reduces to

$$\sigma_u^2 = \frac{M_0 - m_0}{m_0 M_0(M_0 - 1)}\left[\sum_{i=1}^{s} M_i(u_{i\cdot} - U..)^2 + \sum_{i=1}^{s}(M_i - 1)\sigma_i^2\right]. \tag{24}$$

In the following it will be convenient to use the symbols $\bar{\sigma}$ and $(\overline{\sigma^2})$ to denote, respectively, the weighted averages

$$\bar{\sigma} = \frac{1}{M_0}\sum_{i=1}^{s} M_i\sigma_i \tag{25}$$

and

$$(\overline{\sigma^2}) = \frac{1}{M_0}\sum_{i=1}^{s} M_i\sigma_i^2. \tag{26}$$

Then (24) can be rewritten as

$$\sigma_u^2 = \frac{M_0 - m_0}{m_0M_0(M_0 - 1)}\left[\sum_{i=1}^{s} M_i(u_i. - U..)^2 + M_0(\overline{\sigma^2}) - \sum_{i=1}^{s}\sigma_i^2\right]. \tag{27}$$

If $c_1 = c_2 = \cdots = c_s$, the formula for σ^2_{opt} simplifies and becomes

$$\sigma^2_{\text{opt}} = \frac{1}{m_0}(\bar{\sigma})^2 - \frac{1}{M_0}(\overline{\sigma^2}). \tag{28}$$

Therefore,

$$\sigma_u^2 - \sigma^2_{\text{opt}} = \frac{M_0 - m_0}{m_0M_0(M_0 - 1)}\sum_{i=1}^{s} M_i(u_i. - U..)^2 + \frac{M_0^2 - m_0}{m_0M_0(M_0 - 1)}(\overline{\sigma^2}) - \frac{1}{m_0}(\bar{\sigma})^2 - \frac{M_0 - m_0}{m_0M_0(M_0 - 1)}\sum_{i=1}^{s}\sigma_i^2. \tag{29}$$

Using formula (9), we may write

$$(\bar{\sigma})^2 = \overline{\sigma^2} - \frac{1}{M_0}\sum_{i=1}^{s} M_i(\sigma_i - \bar{\sigma})^2. \tag{30}$$

On substituting this expression into (29) and rearranging, we obtain

$$\sigma_u^2 - \sigma^2_{\text{opt}} = \frac{M_0 - m_0}{m_0M_0(M_0 - 1)}\sum_{i=1}^{s} M_i(u_i. - U..)^2 + \frac{1}{m_0M_0}\sum_{i=1}^{s} M_i(\sigma_i - \bar{\sigma})^2 + \frac{M_0 - m_0}{m_0M_0(M_0 - 1)}\left[\overline{\sigma^2} - \sum_{i=1}^{s}\sigma_i^2\right]. \tag{31}$$

This is an important formula. It indicates the methods of stratification of the population for which an optimally stratified sample will yield the best results. Also, this same formula indicates how even an optimum stratified sample may fail. The latter circumstance will certainly occur if the stratification of the population is so unlucky that the means of the strata and also the internal variabilities of the strata are all equal. Then we have

$$u_1. = u_2. = \cdots = u_s. = U.. \tag{32}$$

and

$$\sigma_1 = \sigma_2 = \cdots = \sigma_s = \bar{\sigma}. \tag{33}$$

In this case,

$$\sigma_u^2 - \sigma^2_{\text{opt}} = -(s - 1)\frac{M_0 - m_0}{m_0M_0(M_0 - 1)}(\bar{\sigma})^2 \tag{34}$$

and it appears that even the optimally stratified sample will give less precise results than the unrestrictedly random sample. It is true that, in most cases, the value of (34) will be negligible. Also, it is most improbable that, with reasonable effort at good stratification, the equalities (32) and (33) will be satisfied, even approximately.

Formula (31) indicates that for stratification to be successful when compared with unrestrictedly random sampling, it is necessary to make the particular strata as different as possible (i) with respect to the averages u_i. and (ii) with respect to their internal variability as measured by the σ_i^2.

You will notice that, while these results are interesting theoretically, there is considerable difficulty in applying them in practice. The optimum stratification of a sample depends (1) on the sizes of the strata as represented by the numbers M_i of sampling units in the ith stratum; (2) on the internal variability σ_i^2 of the ith stratum and (3) on the average cost c_i per unit of sampling in the ith stratum. Since the choice of the units of sampling is at our disposal, the numbers M_i are likely to be known exactly. This is not so for the values of σ_i and, probably, not so for the values of c_i.

If we knew the numbers σ_i^2, we would probably know the numbers u_i. and then there would be no need of sampling. It follows that in no practical case is there an exact and immediate application of the formulas I have given. However, in this respect our situation is no worse than it is in any other attempt to apply mathematical results in practice. In every case, the theory applies only approximately to the situation studied and the data substituted into the mathematical formulas are not exact values of the variables concerned, but only approximations. Thus, if the exact values of σ_i are not available, there are ways and means to estimate them approximately. One typical situation occurs when a particular kind of survey is repeated year after year. In this case last year's sample may be used to estimate the values of σ_i for the next year's survey. The theory behind this procedure is that, while the average level of a given characteristic changes considerably from one year to another, the internal variability of the particular strata is much less unstable and, in particular, a stratum that appears more variable than the others during one year is also more variable the next.

Undoubtedly, there are cases where no information is available about the internal variability of the strata. Then, the best you can do is to use a part of the funds available to conduct a preliminary inquiry which, incidentally, will be helpful for training enumerators. The size of this preliminary inquiry may be very moderate. Out of each of the strata a small number of units of sampling, say 20, are selected at random and the values of the relevant characteristic U are established. Then these values are used to estimate the within strata variances σ_i^2. Finally, the estimates

of the variances σ_i^2 are used to determine the optimum stratification of the sample to be drawn for the main part of the survey.

The situation with the cost c_i per unit of sampling is similar. The exact values of the numbers c_i must remain unknown until the books are closed on the survey, but more or less accurate estimates are not difficult to obtain, especially if the same population has been sampled repeatedly. A closer analysis of previous surveys will probably show that within a given stratum the total cost of sampling increases somewhat more slowly than in direct proportion to the number of units selected for the sample. If this is so, the institution concerned will do well to establish for each of the strata a schedule of the following kind: if the number of units sampled is between 20 and 30 say, then the cost per unit will be approximately so much; if the number of units sampled is between 30 and 40, then the cost per unit will be something else; etc. Figures of this kind could then be used to produce tentative values at first and improved values later of the m_i^* according to formula (17).

It is useful to plan the work so that, by the end of the survey, not only the preliminary sampling but also all the data collected could be used to obtain better estimates of the c_i. Additional computation would then show whether or not the stratification of the sample which was actually made was far from optimum, how much accuracy was lost and whether or not it was worthwhile to try to improve on proportional sampling. Naturally, if data are available, such computations should be made before determining the scheme of sampling.

QUESTION BY MR. STOCK: If you were measuring a number of characteristics, to which one would you tie the m_i?

ANSWER: I welcome this question. It is true that a sizeable inquiry is never planned in order to determine a single mean. On the contrary, we are always interested in a number of characteristics of the population studied and we must make a choice between them. In some cases there may be a characteristic of the population which is overwhelmingly more important than the others and then the choice is easy. In other cases there is a group of several characteristics about equally important, and then the situation is more complicated and may be satisfactorily resolved only after some study. Let me illustrate this point on a particular survey conducted by the Institute for Social Problems in which I took part and which brought me into contact with problems of sampling human populations.

The survey was undertaken in connection with a reform of the Polish system of social insurance and was meant to provide a basis for determining the contributions payable by workers and by employers. For this purpose it was necessary to estimate the total number of workers subject

to insurance, their age distribution, family status, etc. All this information was available in the data of the 1931 general census of Poland. Unfortunately, however, a complete tabulation of the census was not expected for some years to come and the actuarial computations had to be based on a sample taken from the original census records. Thus we were faced with the problem of sampling, not the population of living persons, but the census data.

The whole of Poland was divided into 26 strata. The subdivision was made taking into account both the convenience of sampling and the general principle that the strata internally should be as uniform as possible. The way in which the data were stored enforced the adoption of the enumeration district as the unit of sampling.

Although we intended to study many characteristics of the population, we agreed to consider the following six as the most important:

x = total sum at risk connected with the sickness insurance of employed males, aged 20–64.
y = total number of employed males.
z = total number of employed males, aged 20–64.
u = total number of employed males, aged above 64.
v = total number of employed females.
w = total number of insurable population.

Since no precise information was available beforehand about the internal variability of the individual strata, it was necessary to resort to a preliminary inquiry. Table I, compiled from the data in my Polish publication quoted, gives each of the strata, the numbers M_i of elements of sampling and the estimates of the numbers σ_i computed for each of the six characteristics x, y, z, u, v, w.

When following the columns of the estimated standard deviations within the individual strata, you will hardly fail to notice that the standard deviations of the six characteristics are positively correlated. Consider, for example, the last four strata. Although no strict regularity exists, it is obvious that frequently a stratum greatly variable with respect to one characteristic is also variable with respect to the others. This empirical fact has a theoretical explanation and is connected with the circumstance that the characteristics of the particular units of sampling are usually correlated. This correlation may be positive or negative, but the resulting correlation between the corresponding σ_i is always positive. The correlation coefficients between the estimates of σ_i for w on the one hand and for x, y, z, u, v on the other, are given in Table II.

Thus if we stratify the sample so that optimum conditions for one of the characteristics is approached, then as a result of this correlation, we are

TABLE I

Sizes of strata and their internal variability with respect to six important characteristics

Stratum No. i	Size of stratum M_i	Estimates of internal variability of strata in terms of the σ_i					
		x	y	z	u	v	w
1	2,041	.83	16	14	1.1	5.9	52
2	1,185	.95	20	17	2.8	8.2	50
3	3,032	1.66	25	23	2.5	9.6	81
4	371	.97	11	10	1.6	8.2	37
5	249	1.03	14	12	1.6	9.4	36
6	681	.84	19	16	1.4	13.9	68
7	3,432	.48	16	15	1.0	3.2	54
8	801	.79	17	16	.7	5.4	45
9	2,196	1.16	21	18	2.8	12.9	66
10	4,079	1.62	23	21	1.0	9.8	63
11	2,952	.49	9	7	3.0	9.3	20
12	1,123	.83	10	9	.4	3.5	16
13	1,516	1.73	15	12	2.5	12.2	28
14	1,990	1.04	21	18	1.6	20.5	88
15	998	1.88	21	19	2.1	7.6	72
16	762	.62	10	8	1.5	10.1	21
17	2,867	.57	11	10	1.2	9.3	34
18	443	.77	11	10	1.9	16.1	26
19	2,385	1.37	8	11	1.3	8.0	34
20	4,326	.51	12	10	1.2	8.7	27
21	2,985	.36	7	7	1.4	8.8	38
22	1,243	.65	12	9	1.4	9.3	76
23	29,885	.48	8	7	.5	5.3	28
24	18,636	.16	8	6	.9	2.2	21
25	10,906	1.29	35	31	1.0	9.2	86
26	22,299	.27	6	4	.3	6.5	20
Total	123,383	....	..	..	...	....	..

TABLE II

Coefficients of correlation between the estimates of σ_i for w and those for x, y, z, u, v

Characteristic	x	y	z	u	v
Correlation coefficient	.512	.807	.799	.212	.363

likely to do reasonably well for the others. In this particular inquiry of the Institute for Social Problems, it was considered that the total number of insurable workers was the most important characteristic and the stratification of the sample was adjusted in accordance with the variability of w_i.

Table I illustrates other interesting details that occurred in the Polish inquiry and may occur in others. It will be seen from the column of M_i that the sizes of the individual strata varied between very broad limits. In fact, the smallest stratum contains only 249 units of sampling, while the largest contains almost 30 thousand! As I have already mentioned, great care was taken in establishing the strata to use existing information (unfortunately, this information was predominantly qualitative in character, without actual figures) in order to obtain strata internally as uniform as possible. Thus, whenever a large block of the country was left undivided as a single stratum, it was because of the prevailing belief that the block was very uniform. In a number of cases this guess proved successful. For example, the three very large strata, Nos. 23, 24 and 26, are rather uniform with respect to all six characteristics studied. However, whenever guesswork and intuition underlie our actions, surprises are unavoidable and stratum No. 25 provided something like a shock. When the preliminary sample of 15 units indicated such great variability in No. 25, the committee in charge of sampling was inclined to ascribe this occurrence to a random sampling error and to disbelieve the figures obtained. Accordingly, the preliminary sample from stratum 25 was raised to 34 units. Naturally, the new estimate of σ^2_{25} differed from the first, but the conclusion as to the internal variability of this stratum remained unchanged.

In this particular case no harm was done by enlarging the preliminary sample because, in order to complete the sampling, we had to select from the same stratum an additional 300 units. However, the reverse situation with strata 4, 5, and 8 did cause a certain loss in the precision of the final sample. With respect to these strata, small in size, it was believed that, owing to their industrialized character, they would be internally rather heterogeneous. When the preliminary samples contradicted this expectation, the samples were markedly increased with no essential change in the final conclusion. As a result, the estimated values of the m_i^* for the three strata were 5, 3 and 13, respectively. However, in the preliminary sampling we had already selected 33 sampling units from stratum 4, 21 units from stratum 5 and 61 units from stratum 8. Thus the preliminary inquiry oversampled a number of strata and, in consequence, since the funds for sampling were strictly limited, undersampling of the other strata was unavoidable. As a result, after the preliminary inquiry had been completed and a substantial part of the funds had been spent, we were faced with (seemingly) the new problem of how to apportion the balance of the money

among the undersampled strata so as to attain the greatest accuracy of results.

This problem is only seemingly new and is immediately reduced to the use of the same formulae (17). It is obvious that no more sampling was needed from strata which were already oversampled. Thus, the problem of the best use of the money reduced itself to minimizing the total of those terms in formula (7) which referred to the undersampled strata. Naturally, this had to be done with the use of a new value of m_0, equal to the initial value minus the number of units of sampling already selected from the oversampled strata.

QUESTION BY DR. SIDNEY WILCOX: If you had been advising the Italian census people, what specific advice would you have given?

ANSWER: I would have advised them to consider their *circondarî* not as units of sampling but as strata. These strata should have been subdivided into units of sampling as small as the character of the material permitted —parishes, streets, single houses, whatever was possible. As a matter of fact, I remember seeing a footnote in Gini and Galvani's paper in which they themselves suggest that probably their results would have been more satisfactory if, instead of sampling *circondarî,* they had sampled parishes. In this, of course, they are perfectly correct.

There is a special difficulty in carrying out an inquiry based on a random sample, which seems to be worth mentioning. This is psychological in nature. Generally we do not rely on random sampling. Intuitively, we are inclined to think that it is not wise to rely on chance if there is any knowledge available to guide our steps. I have seen many instances where a feeling similar to this has made it difficult to reach a decision on how an inquiry should be carried out. I remember very well the doubts that I myself had. "That's all right in theory," I thought, "but how would this random sampling work in practice?" Then a great discovery satisfied me how to make up my mind; and since that discovery has worked well with other people, I shall mention it to you. It consists in a simple rule: *try and see.* As far as our intuitive feeling against some theoretical result is concerned, there is nothing like an experiment. In the case of a planned inquiry by sampling, and the question of how to sample, I would take some 1000 sheets from the data, consider them as a sampled population and perform on them in detail all the steps of the several alternative methods of sampling that are contemplated. But I must add two warnings.

(a) The population in this experimental sampling, like the populations we study in practice, must be sufficiently heterogeneous.

(b) The size of the random sample you draw in experimental study must contain a sufficient number of units, say 80 or 100.

I am certain that a few trials of this sort will appeal to your intuition and will give you a comfortable feeling of safety in random sampling, in spite of the fact that in sampling randomly you sometimes ignore knowledge of certain details. But you must remember that in following the indications of the theory you make use of some other kind of knowledge, that of mathematical statistics.

QUESTION BY DR. SIDNEY WILCOX: I would like to ask a question that is somewhat related to this matter of drawing the sample. It is fairly common practice to take a list of the elements of sampling and to start with one that is selected by some device or other and then to take every tenth or twentieth on down the list in order to form the sample. This plan is often used instead of setting up a system of random numbers or drawing numbers at random and then selecting the sample according to the model or game of chance. Are there any advantages or disadvantages that one should bear in mind when making use of the device of taking every tenth name on the list, every tenth family, house or district?

ANSWER: I think there is a definite advantage in using a mechanical process of random sampling throughout; that is to say, not taking every tenth unit as listed. Sometimes nothing will be improved and then your tenth or twentieth house will be as good. But there is the possibility, especially in new and properly planned towns, that if you take every twentieth or fifteenth house, you will be synchronized with something very essential in the town itself. I know of one small inquiry where they took a sample of houses in a few villages. As the houses were numbered, they decided to take every fifth or every tenth, and hoped to obtain a very good sample. But what they obtained was something very surprising. After going back to the sampled villages, they found that house No. 1 was always the one belonging to the squire and this disturbed the sample. In new towns it is likely that every block will have the same number of houses. Therefore, if you take every fifth house, you may either omit corners or systematically include all of them, and thus you may introduce a considerable bias in the sample.

It is essential to be clear about the exact nature of the procedure suggested. The process is this. We take the first ten units of sampling listed and select one of them at random. Let x be its order number. Then to form the sample we take the units numbered x, $x + 10$, $x + 20$, $\cdots$, etc. It will be seen that this procedure is equivalent to dividing the population to be sampled into 10 parts,

1st part, sampling units No. 1, 11, 21, 31, $\cdots$
2nd " " " " 2, 12, 22, 32, $\cdots$

3rd part, sampling units No. 3, 13, 23, 33, ···

.

.

.

10th " " " " 10, 20, 30, 40, ···.

Then we treat these parts as units of sampling and take *only one of them* to form a sample.

Obviously, if we proceed in this way we do not rely on the theory of probability but on good luck, with the hope that the ten parts into which the whole population is divided are very similar. This will frequently be the case, but there are obvious dangers. I recommend that one rely on chance as governed by the empirical law of big numbers, but I do not recommend that one rely on good luck.

As a matter of fact, there are no special difficulties in sampling randomly. There is a very useful little book of Tippett's *Random Sampling Numbers* [8] which may be recommended for the purpose. If your sampling units are listed and numbered in order, to take a random sample of them, you simply open the book and read in turn a sufficient amount of numbers. Whenever the same number appears twice, you simply ignore it. Also, you ignore all numbers exceeding the total of your sampling units.

QUESTION BY DR. LANG: I do not see how this system can be applied to names that are listed alphabetically.

ANSWER: Before using Tippett's *Random Sampling Numbers* you will have to number all your names.

In regard to the question just discussed, it may be useful to mention that in many cases every tenth item will give as good a sample as the

[8] L. H. C. Tippett: "Random Sampling Numbers," *Tracts for Computers,* No. XV, Cambridge University Press, 1927, viii + 26 pp.

See also two newer tables of random numbers: R. A. Fisher and Frank Yates: *Statistical Tables for Biological, Agricultural and Medical Research,* Oliver and Boyd, London, 1938, 90 pp. M. G. Kendall and B. Babington Smith: "Tables of Random Sampling Numbers," *Tracts for Computers,* No. XXIV, Cambridge University Press, 1939, x + 60 pp.

In recent times, with the advent of high speed calculators capable of producing rapidly great quantities of "random numbers," and with the increased use of punch card machines, tables of random numbers have given way to sets of random numbers punched on cards. A set of a million or so random numbers on cards should be considered a regular part of the equipment of every modern institution engaged in sampling surveys.

The words "random numbers" are placed in quotation marks. It is hoped that the reader of this book will realize that, strictly speaking, no such thing as a "random number" or a "set of random numbers" can actually exist. What can, and actually does exist is a method of producing numbers imitating successfully the concept of independent sampling from a uniformly distributed population.

application of Tippett's numbers. Other methods may be used also. It is very difficult to give a general rule for distinguishing between reasonable precautions to insure randomness and attempts to "split hairs." Here the research worker must acquire experience and use his own judgment. It must be emphasized, however, that the use of random numbers does not present any difficulty, and that their use puts you on the safe side.

QUESTION BY MR. KANTOR: Suppose that you have to sample the workers in various industries in several states or other geographical areas. You do not have any record of the unemployed, and you want a sample that will give you the percentage of unemployed in each industry for each of the areas. The reason for the different areas is that there may be economic factors that affect the unemployment rate in an area where there is a small part of the industry as contrasted with the area where there is a major center of it or where there is diversified or unified industry. How can one go about getting a sample that would give results equally accurate for each industry within each district?

ANSWER: There is no particular difficulty in approaching the ideal of equally accurate estimates for different areas concerning the same industry but it may be impossible to attain in addition to this a similar equality in accuracy for all industries. The situation you describe is more complicated than the ones we have considered. The different areas you mention, let us call them partial populations, must be considered separately.

In particular, each partial population must be stratified. If the internal variability of each stratum is known or can be estimated, the application of formula (17) will determine the optimum stratification of the sample to be taken from each partial population. Then the optimum variance of the sample mean will become a function of the total number of elements which will be selected from any given partial population. The problem of allocating the available funds to particular partial populations so as to insure the same precision for each will reduce to something very similar to that described above and I am sure will present no new difficulties.

QUESTION BY MR. KANTOR: In attempting to get an estimate of the variability that you are going to use in deciding what proportion to draw, you will have to take a test count in each of your areas; you have the count scattered over a number of characteristics; it is no longer one characteristic that you measure. You would have to get a test drawing for a number of industries in each of your areas and then compute actual unemployment rates. Isn't that the only way in which you can proceed with many industries? It seems to me that you have to take a full count.

ANSWER: I do not think so. The preliminary inquiry designed to estimate the variability of the strata may be very small in size. Dr. Sukhatme

investigated this question [9] and found that 20–30 units of sampling out of each stratum would be plenty. Indeed he suggests as few as 15. Also it is not necessary to make a separate preliminary inquiry for each industry. You make one such inquiry for an area and use it to estimate $\sigma_j(i)$ for each of the industries in turn. Then, substitute your estimate of the true $\sigma_j(i)$ in formula (17), separately for each industry. You will see that this formula will give more or less similar results for all industries. Alternatively, you may adjust the proportions of sampling to some single character treated as basic. I would choose for this the total number of workers within the sampling unit since it is likely to be highly correlated with the numbers of unemployed.

QUESTION BY MR. KANTOR: In industry, we find that there are very great differences in the proportion of unemployed, depending on the production rate of the industry to which the workers are attached. During a depression, the production of goods for use in further production declines very rapidly, but the production of articles made for general consumption declines only slightly; an area devoted principally to the former type of production will have very high unemployment and an area devoted largely to the latter type of production will have small unemployment. Is this the variability that we can test by drawing a small preliminary sample?

ANSWER: The variability of which you speak does not cause any trouble since this is a variability *between* strata or perhaps *between* partial populations. I presume that the distribution of industries over the country is more or less known and that, when stratifying, you will be able to distinguish areas differing in the general character of the prevailing industries. If you look closely into my formulas, you will notice that they depend upon the variability *within* the partial population and, more particularly, *within* the strata. Denote by w the number of workers within a unit of sampling, and by x the number unemployed. If you take one particular stratum and study the units of sampling, you may find a picture something like this:

Values of w	100	150	35	10	200	etc.
Values of x	10	13	1	2	25	

and you will have no difficulty in noticing that x and w are correlated. Because of this correlation, the stratification of the sample which is optimum for w will be reasonably good for x. Something of this sort actually happened in an inquiry in Poland.

[9] P. V. Sukhatme: "Contribution to the theory of the representative method." *Jr. Roy. Soc. Stat. Supplement,* Vol. 2 (1935), pp. 253–268.

The plan to take one basic character as a unit has the advantage that the preliminary inquiry may be very inexpensive and yet satisfactory. The enumerators could be asked to establish only the number of workers inhabiting the units of sampling, a task which takes but very little time and effort. But also this procedure has the definite disadvantage that, if you work with the basic character alone, the data collected during the preliminary inquiry cannot be included in the main one. Therefore, probably I would carry out the preliminary exactly as the main one is to be made, the only difference being one of size. I would estimate σ_i separately for each industry and substitute it into formula (17). Then I would see what happens and what would be the accuracies of the average that I would obtain by this or that system of stratifying the sample.

QUESTION BY MR. MILTON FRIEDMAN: In many cases the set of characteristics that it is desired to study includes some about which information can be obtained with relative ease and others about which information can be secured only through long and expensive interviews. In such cases it may be advisable to secure information on the first set of characteristics from a large random sample. This information may then be used to select a smaller stratified sample from which the second type of data can be secured. From the random sample would also be obtained weights to be used in combining the data from the various strata of the stratified sample.

Thus, in the Study of Consumer Purchases, which is now being conducted under the auspices of the National Resources Committee, the Bureau of Labor Statistics, and the Bureau of Home Economics, the primary aim is to secure information on family expenditures. The sample from which such data are secured is, however, stratified with respect to income (as well as other characteristics). At the same time, there are no data on the relative frequencies of the different income classes. As a consequence, it was necessary to obtain information on income from a random sample of families in order to secure the weights for combining the data from the stratified sample. In view of the extremely high costs involved in securing the data on expenditures, and of the relatively low costs of securing the data on incomes, it was decided to make the random sample from which income information was obtained very much larger than the stratified sample giving information on expenditures.

The question I should like to ask is whether or not any work has been done that would indicate the optimum relative size of the two samples on the assumption that the relative costs and the relevant standard deviations are known.

ANSWER: As far as I know, nothing has been done on the specific question you raise. I take it, however, that in such a case it would be necessary to conduct two preliminary inquiries, one designed to determine the relative

frequency of the different income classes, and the other to determine the standard deviations for the item in which you are particularly interested, for the different strata. The second preliminary investigation, as I have already indicated, would need to cover only a relatively small number of cases.

QUESTION (Mr. Friedman's question restated by Dr. Sidney Wilcox): For at least part of the work one step was taken in trying to get a random sample using every nth card, starting not with the first card but with a card which itself would be the result of accident. This was the process of finding out for a given city what proportion of the people are wage earners and clerical workers and what proportion are at one or another income level. This was an inexpensive survey. Then a long laborious process had to be followed in finding out in detail how they spent their money. The number of families responding to the more elaborate questionnaire might have no very close relationship to the number of families in the particular type of occupational activity or income level. And so the question of weights comes up. What should be the relative number that should be secured on the random basis? Should we take every tenth family or, knowing in advance approximately the costs of the operations and therefore how many schedules we are going to be able to get on the expenditure basis, how heavy a sample should we have taken on the random basis? What is the relative size of the random sample? Of the larger sample to the smaller?

ANSWER: I repeat, as far as I am aware, the question asked has not been considered; but it is so interesting that I shall be glad to see whether or not it can be answered by some simple method. If I succeed, I will certainly try to publish the results.

Part 2. Theory of Friedman-Wilcox Method of Sampling

(This section is a textual reproduction of the article, "Contribution to the Theory of Sampling Human Populations," by the present author, originally published in the *Journal of the American Statistical Association,* Vol. 33 (1938), pp. 101–116. The author is deeply indebted to the Editors of the Journal for their kind permission to reproduce the paper.)

1. INTRODUCTION

At a Conference on Sampling Human Populations held last April at the Department of Agriculture Graduate School in Washington, a problem was presented by Mr. Milton Friedman and Dr. Sidney Wilcox for which I could not offer a solution at the time. Since it seemed to be important and of general interest, I have considered it in some detail. The purpose of this paper is to present the results I have obtained.

2. STATEMENT OF THE PROBLEM

I shall start by describing the problem in much the same form as it was stated to me, without using any mathematical symbols. Then I shall formulate it in mathematical terms. The reader who does not wish to follow the mathematical processes may skip from equation (8) to the results and examples beginning with equation (52) on page 138.

A field survey is to be undertaken to determine the average value of some character of a population, for example, the amount of money which families spend for food in a population of families residing in a certain district. The collection of these data requires long interviews by specially trained enumerators and, hence, the cost per family is quite high. Since the total cost of the survey must be held within the amount appropriated for it, the data must be secured from a small sample of the population. In view of the great variability of the character, the sample appears to be too small to yield an estimate of the desired degree of accuracy.

Now the character is correlated with a second character which can be determined much more readily and at a low cost per family. Since a very accurate estimate of the second character can be secured at relatively small expense, and since for any given value of it, the variation of the original character will be smaller than it is in the whole population, a more accurate estimate of the original character may be obtained for the same total expenditure by arranging the sampling of the population in two steps. The first step is to secure data, for the second character only, from a relatively large random sample of the population in order to obtain an accurate estimate of the distribution of this character. The second step is to divide this sample, as in stratified sampling, into classes or strata according to the value of the second character and to draw at random from each of the strata a small sample for the costly intensive interviewing necessary to secure data regarding the first character.

An estimate of the first character based on these samples may be more accurate than one based on an equally expensive sample drawn at random without stratification. The question is to determine for a given expenditure, the sizes of the initial sample and the subsequent samples which yield the most accurate estimate of the first character.

Let us now enter into the details and introduce the necessary notation. Denote by π the population studied and by X the character of its individuals the average of which, say $\overline{X}$, is to be estimated. This is the character the collection of data on which is costly. Next let Y denote the second character, on which the collection of data is cheap, and which is assumed to be correlated with X. The range of variation of Y in π being more or less known, we shall divide it into s intervals, say

$$\text{from } Y_0 \text{ to } Y_1, \text{ from } Y_1 \text{ to } Y_2, \cdots, \text{ and from } Y_{s-1} \text{ to } Y_s. \qquad (1)$$

Denote by π_i the part of the population π composed of the individuals for which

$$Y_{i-1} \leqq Y < Y_i, \qquad (i = 1, 2, \cdots, s); \tag{2}$$

π_i will be called the ith stratum of the population π. Denote further by

$$p_1, p_2, \cdots, p_s \tag{3}$$

the proportions of the individuals of π belonging to the strata $\pi_1, \pi_2, \cdots, \pi_s$ respectively.

In the following we shall have to consider three different processes of sampling which it is important to distinguish. The first two form the method described by Mr. Friedman and Dr. Wilcox, which I shall further describe as the method of double sampling. The third will serve as a standard of comparison of the accuracy of the method of double sampling. In order to avoid any misunderstanding let us describe all three in detail.

The method of double sampling consists of the following steps:

(i) Out of the population π we select at random N individuals and ascertain for them the values of the character Y. This sample will be denoted by S_1. The sample S_1 is meant to estimate the proportions p_i.

(ii) Now we proceed to sample the strata π_i and this is the second of the sampling processes mentioned. Out of each stratum π_i we select at random m_i individuals which form a sample to be denoted by $S_{2,i}$ and ascertain for each of these individuals, the value of the character X. The samples $S_{2,i}$ serve to estimate the mean value of X in each of the strata π_i. These estimates and the estimates of the proportions (3) obtained previously from the sample S_1, permit us to estimate the grand mean $\overline{X}$.

The combination of (i) and (ii) forms the method of double sampling. Denote by m_0 the sum of the sizes m_i of all the samples $S_{2,i}$, so that

$$m_0 = \sum_{i=1}^{s} m_i \tag{4}$$

and by A and B the costs of ascertaining for one individual the value of X and that of Y respectively. Finally, let C denote the total amount of money available for the collection of data. Then the numbers m_0 and N must be subject to the restriction

$$Am_0 + BN = C. \tag{5}$$

We shall consider what values of m_i, m_0 and N, satisfying conditions (4) and (5), yield the greatest accuracy in estimating the mean value of X by the method of double sampling. This accuracy will then be compared with that attainable in the ordinary way, that is, without the application of the method of double sampling. For this purpose we shall consider a third sampling process by which all the funds C available are spent on selecting at random a

number, say M, of the elements of π and in ascertaining for each of them the value of X. Denote this third sample by S_0. Its size will have to be $M = C/A$. In order to get an idea of the utility of the method of double sampling we shall compare its accuracy with that of the ordinary mean value of X calculated from the sample S_0.

3. FIRST METHOD OF APPROACH

In the present paper [1] we shall make no assumption as to the character of the regression of X on Y in the population π. Denote by $X_1, X_2, \cdots, X_s$, the mean values of X in each of the strata. It follows that the grand mean of X which is to be estimated is

$$\overline{X} = \sum_{i=1}^{s} p_i X_i. \tag{6}$$

Further denote by σ_i the standard deviation of X within the ith stratum.

Denote by n_i the number of individuals drawn in the first sample S_1 which fall within the ith stratum and introduce

$$r_i = \frac{n_i}{N}. \tag{7}$$

Let x_{ij} denote the value of X of the jth individual drawn from the ith stratum to form the sample $S_{2,i}$. Put

$$x_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}. \tag{8}$$

We shall start by considering what function F_1 of the observations, namely, of the numbers (7) and of

$$x_{i1}, x_{i2}, \cdots, x_{im_i} \qquad \text{for } i = 1, 2, \cdots, s \tag{9}$$

would be suitable as an estimate of (6). We shall limit our considerations to homogeneous functions of second order, of the form

$$F_1 = \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{m_j} \lambda_{ijk} r_i x_{jk} \tag{10}$$

where λ_{ijk} is a constant coefficient. Out of all such functions we shall select and term the best unbiased estimate of $\overline{X}$, the one which has the following properties:

(i) The mathematical expectation of F_1 is identically equal to $\overline{X}$.

(ii) The variance of F_1 is smaller than that of any other function of the form (10) having the property (i).

[1] The same problem, under the assumption that the regression of X on Y has a certain known form, forming the second method of approach, will be considered in a later paper.

Denoting by $\mathcal{E}(u)$ the mathematical expectation of any variable u, we may rewrite (i) in the following form.

$$\mathcal{E}(F_1) \equiv \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{m_j} \lambda_{ijk} \mathcal{E}(r_i x_{jk}) \equiv \sum_{i=1}^{s} p_i X_i. \tag{11}$$

When calculating expectations, we shall use the assumption that the population π and all of its strata are so large compared to the sample drawn that the particular drawings can be considered as mutually independent. We shall notice further that, in spite of the fact that the samples $S_{2,i}$ probably will be drawn out of the sample S_1 and not directly from the strata π_i, the variable x_{jk} is independent of r_i. This follows from the circumstance that when we draw the first sample S_1, we do so without any consideration of the values of X. It follows that

$$\mathcal{E}(r_i x_{jk}) = \mathcal{E}(r_i)\mathcal{E}(x_{jk}) = p_i X_j. \tag{12}$$

Substituting (12) in (11) and rearranging, we have

$$\sum_{i=1}^{s} p_i \left(\sum_{j=1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} - X_i \right) \equiv 0. \tag{13}$$

The necessary and sufficient condition for this equality to hold good identically, that is to say, whatever the unknown proportions $p_1, p_2, \cdots, p_s$ may be, is that the coefficients of the p_i vanish, i.e.,

$$\sum_{j=1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} - X_i \equiv 0 \qquad \text{for } i = 1, 2, \cdots, s. \tag{14}$$

As we do not know the values of the X_j, these equalities should again hold good identically, that is to say, whatever the values of the X_j. The equation (14) can be rewritten in the form

$$\sum_{j=1}^{i-1} X_j \sum_{k=1}^{m_j} \lambda_{ijk} + X_i \left(\sum_{k=1}^{m_i} \lambda_{iik} - 1 \right) + \sum_{j=i+1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} \equiv 0 \tag{15}$$

and its identical fulfillment is easily seen to require that

$$\sum_{k=1}^{m_j} \lambda_{ijk} = 0 \qquad \text{for any } j \neq i;\ i, j = 1, 2, \cdots, s$$

and (16)

$$\sum_{k=1}^{m_i} \lambda_{iik} = 1 \qquad \text{for } i = 1, 2, \cdots, s.$$

Equations (16) express the necessary and sufficient conditions for the function F_1 to be unbiased, considered as an estimate of $\overline{X}$. Obviously, there is an

infinite number of systems of coefficients λ_{ijk} satisfying (16) and therefore an infinity of unbiased estimates of $\bar{X}$ of the form (10). We shall now determine the one that we agreed to call the "best," i.e., that which has the smallest possible variance. Let us assume that the values of the λ_{ijk} are fixed somehow satisfying the conditions (16) and calculate the variance of F_1. Denoting it by V_1 we shall have, owing to (6)

$$\begin{aligned} V_1 &= \mathcal{E}(F_1 - \bar{X})^2 \\ &= \mathcal{E}\left\{\sum_{i=1}^{s}(r_i\xi_i - p_iX_i)\right\}^2 \end{aligned} \tag{17}$$

where

$$\xi_i = \sum_{j=1}^{s}\sum_{k=1}^{m_j}\lambda_{ijk}x_{jk} \tag{18}$$

is again independent of r_i. We have further

$$V_1 = \sum_{i=1}^{s}\mathcal{E}\{(r_i\xi_i - p_iX_i)^2\} + 2\sum_{i=1}^{s-1}\sum_{h=i+1}^{s}\mathcal{E}\{(r_i\xi_i - p_iX_i)(r_h\xi_h - p_hX_h)\}. \tag{19}$$

But

$$\begin{aligned} \mathcal{E}(r_i\xi_i - p_iX_i)^2 &= \mathcal{E}[\{(r_i - p_i)\xi_i + p_i(\xi_i - X_i)\}^2] \\ &= \mathcal{E}\{(r_i - p_i)^2\xi_i^2\} + 2p_i\mathcal{E}\{(r_i - p_i)(\xi_i^2 - X_i\xi_i)\} \\ &\quad + p_i^2\mathcal{E}\{(\xi_i - X_i)^2\} \\ &= \mathcal{E}\{(r_i - p_i)^2\}\mathcal{E}(\xi_i^2) + p_i^2\mathcal{E}\{(\xi_i - X_i)^2\} \end{aligned} \tag{20}$$

owing to the independence of ξ_i and r_i and to the fact that $\mathcal{E}(r_i) = p_i$. Now it is known that

$$\mathcal{E}\{(r_i - p_i)^2\} = \mathcal{E}(r_i^2) - p_i^2 = \frac{p_iq_i}{N} \tag{21}$$

with $q_i = 1 - p_i$. Since [2]

$$\mathcal{E}\{(\xi_i - X_i)^2\} = \mathcal{E}(\xi_i^2) - X_i^2, \tag{22}$$

to calculate (20) it will be sufficient to calculate $\mathcal{E}\{(\xi_i - X_i)^2\}$ or the variance of ξ_i. Applying the usual formula for the variance of a linear function of independent variables and remembering that the variance of x_{jk} is denoted by σ_j^2, we have

$$\mathcal{E}\{(\xi_i - X_i)^2\} = \sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda^2_{ijk}. \tag{23}$$

[2] Owing to (16) the expectation of ξ_i is obviously equal to X_i.

It follows that

$$\mathcal{E}\{(r_i\xi_i - p_iX_i)^2\} = \frac{p_iq_i}{N}\left(\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda^2_{ijk} + X_i^2\right) + p_i^2\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda^2_{ijk}. \tag{24}$$

We may now go on and calculate the expectation of the other type of term in (19). We have

$$\begin{aligned}\mathcal{E}\{(r_i\xi_i - p_iX_i)(r_h\xi_h - p_hX_h)\} &= \mathcal{E}(r_ir_h\xi_i\xi_h) - p_ip_hX_iX_h \\ &= \mathcal{E}(r_ir_h)\mathcal{E}(\xi_i\xi_h) - p_ip_hX_iX_h \end{aligned}\tag{25}$$

again owing to the independence of ξ_i and r_i. It is known that

$$\mathcal{E}(r_ir_h) = p_ip_h\left(1 - \frac{1}{N}\right). \tag{26}$$

Further

$$\mathcal{E}(\xi_i\xi_h) = \mathcal{E}\left(\sum_{j=1}^{s}\sum_{k=1}^{m_j}\lambda_{ijk}x_{jk}\sum_{g=1}^{s}\sum_{u=1}^{m_g}\lambda_{hgu}x_{gu}\right). \tag{27}$$

Remembering that

$$\mathcal{E}(x_{jk}) = X_j \quad \text{and} \quad \mathcal{E}(x^2_{jk}) = \sigma_j^2 + X_j^2 \tag{28}$$

and that the x_i are assumed to be mutually independent, we have

$$\mathcal{E}(\xi_i\xi_h) = \sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} + \left(\sum_{j=1}^{s}X_j\sum_{k=1}^{m_j}\lambda_{ijk}\right)\left(\sum_{g=1}^{s}X_g\sum_{u=1}^{m_g}\lambda_{hgu}\right). \tag{29}$$

Until the present moment we have not used the conditions (16) for the unbiased character of the estimate F_1. Therefore the formula for the variance V_1 which we could obtain by substituting (24), (25), (26) and (29) into (19) would be perfectly general. We shall use it in our second method of approach. Now, however, we shall simplify (29) by substituting (16). We have

$$\mathcal{E}(\xi_i\xi_h) = \sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} + X_iX_h. \tag{30}$$

Now

$$\begin{aligned}V_1 = \sum_{i=1}^{s}\left(p_i^2 + \frac{p_iq_i}{N}\right)\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda^2_{ijk} + \sum_{i=1}^{s}\frac{p_iq_i}{N}X_i^2 \\ + \frac{2}{N}\sum_{i=1}^{s-1}\sum_{h=i+1}^{s}p_ip_h\{(N-1)\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} - X_iX_h\}.\end{aligned}\tag{31}$$

Without attempting to simplify this expression at the present stage, let us select the λ_{ijk} so as to minimize (31) while keeping the relations (16) satisfied. For this purpose we will differentiate with respect to λ_{ijk} the expression

$$f = V_1 - 2\sum_{i=1}^{s}\sum_{j=1}^{s}\alpha_{ij}\sum_{k=1}^{m_j}\lambda_{ijk} \tag{32}$$

where the α_{ij} are Lagrange arbitrary multipliers, and equate the derivatives to zero. After some rearrangement, we get the following equation:

$$\frac{1}{N}p_i\sigma_j^2\{\lambda_{ijk} + (N-1)\sum_{h=1}^{s}p_h\lambda_{hjk}\} = \alpha_{ij}. \tag{33}$$

Summing both sides with respect to k from zero to m_j and taking into account (16), we get

$$\begin{aligned} \frac{N-1}{N}p_jp_i\sigma_j^2 &= m_j\alpha_{ij} \qquad \text{for } i \neq j \\ \frac{1}{N}p_{ij}\sigma_j^2\{1 + (N-1)p_j\} &= m_j\alpha_{jj}. \end{aligned} \tag{34}$$

Substituting these results in (33), we obtain

$$\lambda_{ijk} = \frac{N-1}{m_j}p_j - (N-1)\sum_{h=1}^{s}p_h\lambda_{hjk} = \lambda_{\cdot jk} \quad \text{(say)} \tag{35}$$

$$\lambda_{jjk} = \lambda_{\cdot jk} + \frac{1}{m_j}. \tag{36}$$

Substituting in (35) the values of λ_{hjk} thus obtained, we easily get

$$\begin{aligned} \lambda_{ijk} &= \lambda_{\cdot jk} = 0 \qquad \text{for } i \neq j \\ \lambda_{jjk} &= \frac{1}{m_j}. \end{aligned} \tag{37}$$

Substituting these values into (10) we obtain the following expression for the best unbiased estimate of $\bar{X}$:

$$F_1 = \sum_{i=1}^{s}r_ix_i. \tag{38}$$

The formula for the variance, V_1, of F_1 is obtained by substituting (37) in (31)

$$V_1 = \sum_{i=1}^{s}\left\{\left(p_i^2 + \frac{p_iq_i}{N}\right)\frac{\sigma_i^2}{m_i} + \frac{p_iq_i}{N}X_i^2\right\} - \frac{2}{N}\sum_{i=1}^{s-1}\sum_{j=i+1}^{s}p_ip_jX_iX_j \tag{39}$$

which immediately reduces to the following form most convenient for finding the system of values of N and the m_i that assure the greatest accuracy in estimating $\bar{X}$:

$$V_1 = \frac{1}{m_0}\left(\sum_{i=1}^{s} \sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}\right)^2 + \sum_{i=1}^{s} m_i \left(\frac{\sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}}{m_i} - \frac{\sum_{i=1}^{s} \sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}}{m_0}\right)^2 + \frac{1}{N}\sum_{i=1}^{s} p_i (X_i - \bar{X})^2. \quad (40)$$

It is seen that none of the three terms in the right hand side can be negative. There is only one term which depends directly on $m_1, m_2, \cdots, m_s$, namely, the second, the others being dependent on $m_0 = \sum_{i=1}^{s} m_i$ and on N. It follows that once N and m_0 are fixed in one way or another the value of V_1 depends on the m_i and the value they ascribe to the second term. It is easily seen that its minimum value is zero and that this is attained whenever for each value of $i = 1, 2, \cdots, s$

$$m_i = \frac{m_0 \sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}}{\Sigma \sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}}. \quad (41)$$

Owing to the fact that the m_i are integers, this ideal seldom can be attained exactly, but it may be approached as far as possible. We shall further assume that the m_i are selected in closest agreement with (41) and that the second term in (40) is negligible compared with the remaining two.

We must now consider what values of m_0 and N satisfying (5) are likely to give the smallest value to the sum of only two terms in (40), say

$$V_1' = \frac{1}{m_0}\left(\sum_{i=1}^{s} \sigma_i \sqrt{p_i^2 + p_i q_i N^{-1}}\right)^2 + \frac{1}{N}\sum_{i=1}^{s} p_i (X_i - \bar{X})^2. \quad (42)$$

Owing to the complex structure of the first of these terms, an accurate solution of the problem is difficult to attain. However, it is easy to get an approximate solution which will probably in most cases be sufficient.

In most cases, whenever we do not make any special assumption concerning the character of the regression of X on Y, we shall probably classify the population π into only a few strata whence it may be assumed that the proportions p_i will not be very small and consequently $p_i q_i N^{-1}$ will be considerably smaller than any of the p_i^2. If so, then the value of the square root

$$\sqrt{p_i^2 + p_i q_i N^{-1}} \quad (43)$$

will be very much the same as that of p_i. For example, if $p_i = .1$, $q_i = .9$ and $N = 100$, it is .1044 and if the value of Np_i were somewhat larger, the

agreement would be still better. Therefore, instead of trying to minimize (42) we may usefully start by trying to minimize, say

$$V_1'' = \frac{1}{m_0}\left(\sum_{i=1}^{s} p_i\sigma_i\right)^2 + \frac{1}{N}\sum_{i=1}^{s} p_i(X_i - \bar{X})^2,$$

or

$$V_1'' = \frac{a^2}{m_0} + \frac{b^2}{N} \tag{44}$$

for short. Denote by v_1 and v_2 the smallest numbers of selections into the first and the second sample respectively, the total cost of which is the same, so that

$$v_1 B = v_2 A. \tag{45}$$

If m_0' and N' are the integer numbers minimizing (44) and satisfying (5), then any change of these values by taking instead of them either

$$m_0' - v_2 \quad \text{and} \quad N' + v_1$$

or

$$m_0' + v_2 \quad \text{and} \quad N' - v_1 \tag{46}$$

will increase the value of (44). This means that m_0' and N' satisfy the inequalities

$$\frac{a^2}{m_0' + v_2} + \frac{b^2}{N' - v_1} > \frac{a^2}{m_0'} + \frac{b^2}{N'} < \frac{a^2}{m_0' - v_2} + \frac{b^2}{N' + v_1}. \tag{47}$$

These inequalities reduce easily to the following ones

$$\frac{1 - \dfrac{v_2}{m_0'}}{1 + \dfrac{v_1}{N'}} < \frac{a^2 v_2}{m_0'^2}\frac{N'^2}{b^2 v_1} < \frac{1 + \dfrac{v_2}{m_0'}}{1 - \dfrac{v_1}{N'}} \tag{48}$$

showing that in order to minimize (44) while keeping (5) fixed, we have to select m_0 and N as nearly as possible proportionately to $a\sqrt{v_2}$ and $b\sqrt{v_1}$ respectively. Putting for a moment

$$m_0 = N\frac{a}{b}\sqrt{\frac{v_2}{v_1}} \tag{49}$$

and substituting it in (5), we get

$$N = \frac{Cb\sqrt{v_1}}{Aa\sqrt{v_2} + Bb\sqrt{v_1}} \tag{50}$$

which gives

$$m_0 = \frac{Ca\sqrt{v_2}}{Aa\sqrt{v_2} + Bb\sqrt{v_1}}. \tag{51}$$

Using (45) and eliminating v_1 and v_2 we may rewrite (50) and (51) in the final form

$$N = \frac{Cb}{a\sqrt{AB} + bB} \tag{52}$$

$$m_0 = \frac{Ca}{aA + b\sqrt{AB}} \tag{53}$$

where

$$a = \sum_{i=1}^{s} p_i\sigma_i \tag{54}$$

and

$$b^2 = \sum_{i=1}^{s} p_i(X_i - \overline{X})^2. \tag{55}$$

Here we must remember the following circumstances:

(1) that both m_0 and N are integers and therefore formulae (52) and (53) should be calculated to the nearest integer;

(2) that a change in m_0 by one unit must be compensated by a change in N by several units;

(3) that the solutions which would be obtained by taking exact values of (52) and (53) would minimize the value of (44) with a as given in (54), whereas the value of the variance in (42) depends on

$$a_1 = \sum_{i=1}^{s} \sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}} \tag{56}$$

instead of a.

It follows that the integers nearest to (52) and (53) may not necessarily minimize (42), but since the difference between a and a_1 is slight, they may be considered as the first approximations. Frequently these first approximations will also be the accurate values.

In order to find the second approximation, we may calculate a_1 as in (56) substituting N as calculated from (52) and then substitute the value obtained into (53) to get a new value of m_0. This sometimes will indicate the necessity of increasing the original m_0 by unity. However, owing to the fact that both m_0 and N must be integers, the real check of what values do give the minimum is obtained simply by substituting into (42) both the first approximations to m_0 and N and a few neighboring systems of values, e.g., $m_0 - 1$ and $m_0 + 1$ and the corresponding values of N.

4. EXAMPLE I

It may be useful to illustrate the above theory by some simple examples. Assume that there are only three strata, so that $s = 3$. Assume further the following values of the constants involved:

$$\begin{aligned} p_1 &= \tfrac{1}{4}, & p_2 &= \tfrac{2}{4}, & p_3 &= \tfrac{1}{4}, \\ X_1 &= 1, & X_2 &= 3, & X_3 &= 6, \\ \sigma_1 &= 1, & \sigma_2 &= 2, & \sigma_3 &= 4, \\ A &= 4, & B &= 1, & C &= 500. \end{aligned} \tag{57}$$

In order to calculate the values of m_0 and N, we calculate

$$a = 2.25, \tag{58}$$

$$\bar{X} = 3.25, \tag{59}$$

$$b^2 = 3.1875 = (1.7854)^2. \tag{60}$$

It follows that

$$N = 142 \quad \text{(to the nearest integer)} \tag{61}$$

and accordingly

$$m_0 = 89. \tag{62}$$

It will be seen that the necessity of taking m_0 to the nearest integer permits an increase in the value of N to 144, without exceeding the limit of expense, 500 units. Let us now see how $m_0 = 89$ should be distributed between the three strata. Easy calculations give

$$\begin{aligned} \sigma_1\sqrt{p_1^2 + p_1q_1N^{-1}} &= .2526 \\ \sigma_2\sqrt{p_2^2 + p_2q_2N^{-1}} &= 1.0035 \\ \sigma_3\sqrt{p_3^2 + p_3q_3N^{-1}} &= 1.0104 \\ \sum_{i=1}^{3} \sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}} &= 2.2664. \end{aligned} \tag{63}$$

Hence, using (41) and taking the nearest integers, we get

$$m_1 = 10, \qquad m_2 = 39, \qquad m_3 = 40. \tag{64}$$

With this system of the m_i the middle term of formula (40) would have the value

$$\sum_{i=1}^{s} m_i \left(\frac{\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}}{m_i} - \frac{\Sigma\, \sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}}{m_0} \right)^2 = .0000048564. \tag{65}$$

The total value of V_1 in (40) is found to be

$$V_1 = .079855 \tag{66}$$

and it follows that by using (64) the value of the middle term is for all practical purposes negligible. It is interesting to compare this value with the one which could be obtained without adjusting the numbers m_i to the variability and the size of the strata, i.e., without using (41). Putting arbitrarily $m_1 = 29$, $m_2 = m_3 = 30$, we get

$$V_1 = .091927. \tag{67}$$

Comparing this with (66) we see that neglecting to adjust the m_i according to formula (40) results, in this particular example, in an increase of the variance by over 15 percent, which is a considerable and unnecessary loss in accuracy.

This is the situation if we use for m_0 and N the values found as first approximations. Substituting 144 for N in (56) and calculating a_1 and then using this value instead of a to calculate the second approximation of m_0, we get

$$m_0 = 89.6783 \tag{68}$$

which suggests that the best integer values of m_0 and N are $m_0 = 90$ and $N = 140$. However, using them we obtain

$$V_1 = .079866. \tag{69}$$

Again using $m_0 = 88$ and $N = 148$ we get

$$V_1 = .079888 \tag{70}$$

and it appears that the first approximation gives in fact the best possible result, but the actual difference is negligible.

We must now see whether this result, the best that could be obtained by the method of double sampling is actually better than what could be obtained by spending all the money available to collect as much data on X as possible, i.e. by drawing the unrestricted random sample S_0 (see p. 130).

The best linear estimate of $\overline{X}$ calculated from the sample S_0 would be the sample mean $\bar{x}$. Its variance, V_0, is known to be connected with the symbols of this paper by means of the formula

$$V_0 = \frac{1}{M}\left\{\sum_{i=1}^{s} p_i\sigma_i^2 + \sum_{i=1}^{s} p_i(X_i - \overline{X})^2\right\}. \tag{71}$$

It is easy to find that in our example

$$M = \frac{C}{A} = 125 \tag{72}$$

and

$$V_0 = .0755. \tag{73}$$

It follows that in this particular case the method of double sampling, even supplemented by the optimum adjustment of the numbers of sampling, is equivalent to a certain loss of accuracy of the final result. Taking the ratio of the variances (73) and (66)

$$\frac{V_1}{V_0} = 1.058 \tag{74}$$

we see that this loss of accuracy amounts to nearly 6 percent. This unfavorable result is, of course, due to the fact that the differentiation between the strata with respect to the values of X is small compared with the variability of the strata themselves and to the fact that the difference in the cost of obtaining data on X and Y is comparatively small. To illustrate this point let us consider the following examples.

5. EXAMPLE II

Assume that the values of the p_i, X_i and σ_i are exactly as in Example I and put

$$A = 40, \qquad B = 1, \qquad C = 5000 \tag{75}$$

so that the process of obtaining data on Y is now 40 times cheaper than that on X, while the ratio of C/A is the same as formerly. It follows that V_0 in this case will be exactly the same as formerly (73), but the minimum value of V_1 will change. We shall have

$$m_0 = 111, \qquad N = 560 \tag{76}$$

and, assuming that the m_i are fixed according to (41), we get finally

$$V_1 = .05147 \tag{77}$$

and it is seen that this value is exceeded by V_0 by more than 46 percent!

6. EXAMPLE III

Here we shall keep the values of the p_i, the σ_i, and those of A, B and C as in Example I but change the values of the X_i so as to increase the value of b, namely, put

$$X_1 = 1, \qquad X_2 = 6, \qquad X_3 = 11. \tag{78}$$

Then

$$b^2 = 12.5 = (3.53553)^2 \tag{79}$$

and

$$V_0 = .1500. \tag{80}$$

On the other hand, applying the method of double sampling and taking the optimum system of numbers of samplings, viz.,

$$m_1 = 8, \quad m_2 = 31, \quad m_3 = 31, \quad m_0 = 70, \quad N = 220 \quad (81)$$

we get

$$V_1 = .1298, \quad (82)$$

a gain in accuracy in comparison with (80) of about 15 percent.

7. CONCLUSIONS

(i) The examples II and III show that under favorable conditions the method of double sampling is a very powerful tool of statistical research.

(ii) However, the advantages of methods are but rarely universal and in certain cases, as for instance in the above example I, the direct unrestricted sampling may be more efficient than the method of double sampling.

(iii) Without a certain previous knowledge of the properties of the population sampled it is impossible to say which of the two methods will be more efficient.

(iv) It is also impossible to tell in advance what the values of N, m_0, and of the m_i should be to assure the greatest accuracy of the double sampling method.

(v) On the other hand, if certain properties of the sampled population π are known, or can be estimated, then it is possible to estimate the values of m_0 and N and also those of the m_i by which the method of double sampling gives the greatest possible accuracy. The properties of population π needed for this purpose are the values of the p_i, σ_i and X_i. They could be estimated by means of a preliminary inquiry on the lines suggested by me during the conference at the U. S. Department of Agriculture Graduate School and also in my previous publications on sampling human populations.[3] Once approximate values of the p_i, σ_i and X_i are obtained, they should be substituted into formulae (52), (53) and (41) to obtain the approximations of the optimum values of m_0, N and the m_i.

(vi) Before deciding whether to apply the method of double sampling, we should see that the prospects are that it will give better results than the direct unrestricted sampling of values of X.

For this purpose the approximate values of the p_i, σ_i and X_i should be substituted into (40) and (71) to obtain the approximate values of variance V_1 and V_0. The decision to apply the method of double sampling should be

[3] J. Neyman: "An Outline of the Theory and Practice of Representative Method Applied in Social Research." Institute for Social Problems, Warsaw, 1933. Polish with an English Summary.

J. Neyman: "On the Two Different Aspects of the Representative Method." J.R.S.S. 1934, pp. 558–625.

See also P. V. Sukhatme: "Contribution to the Theory of the Representative Method." Supplement to the J.R.S.S., Vol. II, 1935, pp. 253–268.

taken only if the approximate value of V_1 proves to be considerably smaller than that of V_0.

(vii) The steps described in (iii) and (iv) are possible only if some previous knowledge of the population π is available. This may be obtained from various sources: from some previous experience concerning the population π, or from a specially arranged preliminary inquiry. Such a preliminary inquiry consists of drawing from π a relatively small unrestricted random sample of individuals and in ascertaining for all of them the values of both characters under consideration X and Y. The data thus obtained should be used to estimate the p_i, the σ_i and the X_i.

In order to exemplify the kind of previous experience which may be used to plan future inquiries on the lines as indicated in (v) and (vi), I may mention a recent extensive Study of Consumer Purchases, a Federal Works Project administered by the Bureau of Labor Statistics, U. S. Department of Labor and the Bureau of Home Economics, U. S. Department of Agriculture, in cooperation with the National Resources Committee and the Central Statistical Board.[4] This inquiry was carried out by method of double sampling and therefore, in the process of working out the data, both the proportions p_i and the means X_i corresponding to particular strata and to many a character X must have been estimated. Probably the values of σ_i are also available. These figures could be used as pointed out in (v) and (vi) when planning any new inquiry concerning the same characters and the same or some similar population.

Part 3. On a Most Powerful Method of Discovering Statistical Regularities

(This section is based on a talk given before the members of Sigma Xi at a meeting of the Society held in Berkeley, California, April 9, 1947.)

You must have heard the often repeated joke that there are three kinds of lies: the polite lie, the malicious lie and statistics. The subject of my talk tonight will be the kind of statistics that is frequently a lie although, undoubtedly, the authors compiling such statistics do not mean any sort of mischief. For the most part they are well meaning but ignorant of the theory of statistics and they are the victims of their own lack of professional education.

There are many ways of handling perfectly correct data which at first sight seem intuitively sound but which tend to introduce into the data extraneous regularities. These regularities, artificially introduced into the observational material, suggest connections between the various factors

[4] Jour. Am. Stat. Assoc., Vol. XXXI, 1936, p. 135, and Vol. XXXII, 1937, p. 311.

which in fact do not exist. My purpose tonight is to describe one example of such an analysis of statistical data. If you look through the volumes of a statistical-economical or a statistical-sociological journal, you are very likely to find examples of practical applications of this method.

The method in question is so powerful that by means of it one can successfully prove that storks bring babies. Once upon a time an inquisitive friend of mine decided to study this question empirically and thereupon he collected some relevant data. The data are quite comprehensive and refer to 54 different counties. The raw data which he collected are reproduced in Table I.

The data include the number W of women of child-bearing age (second column of Table I, given in units of 10,000), the number S of storks in each county (third column) and, finally, the number B of babies born during a specified period of time (fourth column). In the beginning, my friend had in mind a direct comparison of the numbers S and B. However, it was pointed out to him that such a comparison is not convincing because the counties vary in size and larger counties may be expected to have more women, more babies and also more storks. Thus the variation in the size of the county appeared as a disturbing factor hiding the true relationship between the two quantities S and B.

In order to eliminate the disturbing influence of the size of the county, my friend hit upon the brilliant idea of comparing, not the actual numbers of births and the actual numbers of storks, but the birth rates on the one hand and the "densities of storks" per 10,000 women on the other.

Thus he obtained the quantities X and Y as follows,

$$X = \frac{S}{W} \quad \text{and} \quad Y = \frac{B}{W},$$

and then he tried to compare the two quotients X and Y. Naturally, in questions of this kind you cannot expect an absolute regularity. In particular, you cannot possibly expect that every increase in the quotient X will always be accompanied by a proportional increase in Y. There must be fluctuations and so you will expect to find counties with a large density of storks and a small birth rate and vice versa. The best you can hope for in the way of regularity is that, if you classify all 54 counties according to the density of storks and average the corresponding birth rates, then the averages will show a variation parallel to the variation in the density of storks. To put it professionally: it is inconceivable that the birth rate is a monotonic function of the density of storks and the believer in the proficiency of these birds must be satisfied if he finds a positive correlation.

This was the attitude of my friend and he compiled Table II. I have checked the figures in Table II and so have several other people. We found

TABLE I

Do storks bring babies?—Raw data

County no.	Women in 10,000s	Storks	Babies born	County no.	Women in 10,000s	Storks	Babies born
1	1	2	10	28	4	6	25
2	1	2	15	29	4	6	30
3	1	2	20	30	4	6	35
4	1	3	10	31	4	7	25
5	1	3	15	32	4	7	30
6	1	3	20	33	4	7	35
7	1	4	10	34	4	8	25
8	1	4	15	35	4	8	30
9	1	4	20	36	4	8	35
10	2	4	15	37	5	7	30
11	2	4	20	38	5	7	35
12	2	4	25	39	5	7	40
13	2	5	15	40	5	8	30
14	2	5	20	41	5	8	35
15	2	5	25	42	5	8	40
16	2	6	15	43	5	9	30
17	2	6	20	44	5	9	35
18	2	6	25	45	5	9	40
19	3	5	20	46	6	8	35
20	3	5	25	47	6	8	40
21	3	5	30	48	6	8	45
22	3	6	20	49	6	9	35
23	3	6	25	50	6	9	40
24	3	6	30	51	6	9	45
25	3	7	20	52	6	10	35
26	3	7	25	53	6	10	40
27	3	7	30	54	6	10	45

no mistakes in arithmetic. Furthermore, you will have no difficulty in checking the table yourself. Among the 54 counties studied, there were three in which there were on the average 1.33 storks per 10,000 women of child-bearing age. The average birth rate in these counties was 6.67.

TABLE II

Do storks bring babies?—Analytical presentation

Density of storks per 10,000 women	Number of counties	Average birth rate	Class average
1.33	3	6.67	7.12
1.40	3	7.00	
1.50	6	7.08	
1.60	3	7.00	
1.67	6	7.50	
1.75	3	7.50	9.22
1.80	3	7.00	
2.00	12	10.21	
2.33	3	8.33	11.67
2.50	3	10.00	
3.00	6	12.50	
4.00	3	15.00	

Also, there were three counties with 1.40 as the density of storks and the average birth rate for these was 7.00, and so forth down the column. An inspection of Table II will show that the birth rate, although subject to fluctuations, steadily increases with an increase in the density of storks. This increase becomes even more marked if we divide all the counties into three classes according to the density of storks: densities below 1.7, densities between 1.7 and 2.1, densities above 2.1. The corresponding class averages are given in the last column of the table and show a decisive increase.

My friend's conclusion was that, although there is no evidence of storks actually bringing babies, there is overwhelming evidence that, by some mysterious process, they influence the birth rate! I know that some of you are skeptical and suspect that the original data of Table I were intentionally falsified to produce the astounding result exhibited in Table II. Let me assure you that these suspicions are unfounded. If anything, my friend was extremely lucky in collecting the data. Further, he was certainly very careful in classifying them in Table I so that it is extremely easy to make a complete analysis without performing any arithmetic.

You will notice that all the 54 counties fall into six different groups. It happens that the nine counties forming a group have the same number of women, 10,000 in the first group, 20,000 in the second, etc. Proceeding further, we notice that each group of nine counties falls into three sub-

groups of three counties each. The subgroups are ordered according to the number of storks. Thus, the first group of counties contains three with 2 storks, three with 3 storks and three with 4 storks. The same kind of thing is repeated in all other groups of counties. The counties of the second group must have a larger area than those of the first. They have more women and the number of storks in them varies from 4 to 6, etc. Turning to the columns giving the total number of babies born, we notice constant fluctuations. Thus, within the first group, in the three counties with the same number 2 of storks, the numbers of babies born are 10, 15 and 20. In the next subgroup of three counties there were 3 storks each and, impressed by Table II, we might expect that the numbers of babies, though fluctuating, will show an increase compared with the first subgroup. However, Table I does not display an increase in the number of babies born corresponding to an increase in storks, as long as the number of women remains constant. This is true in the first group of nine counties and it is also true in any other group. So long as we consider a group of counties with *the same number* of women, an increase in the number of storks does not have any effect whatsoever on the number of babies born. We express this technically by saying that the conditional distribution of the number of babies born, given the number of women, is independent of the number of storks. Also, we may say that, given the number of women, the birth rate is independent of the number of storks. This finding appears to be contrary to the intuition of my friend who was much aggrieved, but it coincides with your intuition and my own. Thus, apart from a rather unusual regularity, the figures in Table I do not involve anything unexpected.

How then can one explain the most unexpected features of Table II? Once you start to think about it, the explanation is very easy. The phenomenon was first noticed by Karl Pearson some fifty years ago and was called "spurious correlation."

The variable X, representing the density of storks, is a function of two variables S and W, say

$$X = f_1(S, W) = \frac{S}{W}.$$

Similarly, the birth rate Y is a function of B and W, say

$$Y = f_2(B, W) = \frac{B}{W}.$$

It happens in the present case that the two functions f_1 and f_2 coincide since they are both quotients with W in the denominator. However, the coincidence of the two functions f_1 and f_2 is not essential. The essential

point is that these two functions depend on a common argument W and the fluctuations of W must create simultaneous effects upon the values of X and Y. Since W appears in the denominator of both fractions, any "abnormal" increase in W tends to diminish both X and Y simultaneously. Also, any "abnormal" decrease in W tends to increase both X and Y. As a result, X and Y are positively correlated.

In another case we may be considering two functions, say

$$X = \frac{S}{W} \quad \text{and} \quad Z = BW,$$

where the letters S, B and W stand for some other observable variables. You will easily guess that in this case the presence of W in both X and Z will tend to create a negative correlation between X and Z. This correlation has nothing to do with social or economic factors governing the variation of the three variables but is simply the result of our own arithmetic operations. These are, of course, only intuitive considerations and the exact conclusions require some algebra.

You may be amused by computing the correlation coefficient R between the variables X and Y. This is quite easy if we make certain simplifying assumptions.

We shall assume that

(i) Given W, the variables S and B are independent;

(ii) S and W are correlated and the regression of S on W is linear, say

$$E(S \mid W) = A_0 + A_1 W.$$

Moreover, we shall assume that the conditional variance of S given W, say $\sigma^2_{S \mid W}$, is independent of W.

(iii) B and W are correlated and the regression of B on W is linear, say

$$E(B \mid W) = C_0 + C_1 W.$$

The conditional variance of B given W, say $\sigma^2_{B \mid W}$, is independent of W.

(iv) The expectation and the variance of the reciprocal of W exist. We shall denote them by $1/W_1$ and $\sigma^2_{W^{-1}}$, respectively.

According to the usual definition,

$$R = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}.$$

Thus, in order to compute R, we have to compute the expectations of X, Y, X^2, Y^2 and XY. Easy algebra gives

$$E(X) = E\left(\frac{S}{W}\right) = E\left[\frac{1}{W}E(S \mid W)\right]$$

$$= E\left[\frac{A_0 + A_1 W}{W}\right]$$

$$= \frac{A_0}{W_1} + A_1.$$

Similarly,

$$E(X^2) = E\left[\frac{E(S^2 \mid W)}{W^2}\right]$$

$$= E\left[\frac{\sigma^2{}_{S \mid W} + (A_0 + A_1 W)^2}{W^2}\right]$$

$$= \left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right)(\sigma^2{}_{S \mid W} + A_0{}^2) + \frac{2A_0A_1}{W_1} + A_1{}^2.$$

And it follows that

$$\sigma_X{}^2 = \left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right)\sigma^2{}_{S \mid W} + A_0{}^2\sigma^2{}_{W^{-1}}.$$

Similarly,

$$E(Y) = \frac{C_0}{W_1} + C_1,$$

$$\sigma_Y{}^2 = \left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right)\sigma^2{}_{B \mid W} + C_0{}^2\sigma^2{}_{W^{-1}}$$

and

$$E(XY) = A_0C_0\left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right) + \frac{A_0C_1 + A_1C_0}{W_1} + A_1C_1.$$

Upon substituting these results into the formula for the correlation coefficient, we obtain the final result,

$$R = \frac{A_0C_0\sigma^2{}_{W^{-1}}}{\sqrt{\left\{\left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right)\sigma^2{}_{S \mid W} + A_0{}^2\sigma^2{}_{W^{-1}}\right\}\left\{\left(\sigma^2{}_{W^{-1}} + \frac{1}{W_1{}^2}\right)\sigma^2{}_{B \mid W} + C_0{}^2\sigma^2{}_{W^{-1}}\right\}}}.$$

It follows that the above intuitive considerations are only partly true. In the conditions under which the formula for R was deduced, it is necessary and sufficient for the lack of correlation between X and Y that either $A_0 = 0$ or $C_0 = 0$ or both. If neither of these parameters is zero, then

the correlation R is positive whenever A_0 and C_0 have the same sign and negative otherwise. You will remember that A_0 and C_0 are the intercepts of the regression lines of S on W and of B on W, respectively.

The above analysis of correlation was made under very simplifying assumptions. However, you will have no difficulty in performing it in the more general case when the regressions of S on W and of B on W are represented by any polynomials.

The theory of spurious correlation is not a very difficult matter and, as I have already mentioned, the phenomenon has been known for quite some time. With all due respect to Karl Pearson, I am inclined to alter slightly the label he invented. There is nothing spurious in the correlation between X and Y. When $A_0 \neq 0$ and $C_0 \neq 0$, the correlation between these two variables is quite real. Therefore the term "spurious correlation" seems to miss the point. The real point of the discussion is that the computation of the quotients X and Y is undertaken in order to study the correlation, not between these variables themselves, but between the social, economic or biological factors that these quotients are supposed to represent. It is the method of study that is faulty and, if the adjective "spurious" is to be used at all, it should be applied to the method of studying correlation between factors of primary interest; in the present case, between the number of babies born on the one hand and the number of storks on the other. Only these two factors are of interest. It is suspected that each may be correlated with the third factor, the number of women. Therefore, the appropriate method of study is to compute the partial correlation between B and S with the influence of W eliminated. In proceeding in this fashion, there may be specific difficulties due to the lack of linearity, etc. However, these difficulties can hardly be decreased by using a spurious method.

In spite of the fact that the phenomenon of spurious correlation has been known for half a century, many a practical statistician, as well as the general public, is misled by it from time to time.

Thus we see "proofs" that the density of bars increases the frequency of crimes. This fact is likely to be true but the argument brought in support of the assertion is faulty. In this and similar cases, there is no special harm done to Society. But there are other cases. Not so very long ago, I saw a detailed analysis of various problems of farm management. A considerable amount of money and effort was expended to collect the data. One of the conclusions reached was that, while the primary factor governing the employment of manual labor is the size of the farm, the density of employment increases with the increase of the proportion of the farm land which is arable. Again, this assertion may be true but the argument is faulty and the final tables presented in support of the assertion, quite analogous to Table II, are entirely irrelevant.

There are even more regrettable cases on record in which the spurious method of studying correlation and regression was used and, moreover, was left undetected. Some years ago a scholar was interested in the question of whether or not railway rates were sufficient to cover expenses. His conclusion was that passenger traffic, on the average, barely paid its way and had fallen appreciably short at times while freight traffic, on the average, paid the net of the railway operation. For his analysis he used the data of Class I railroads as given in the annual volumes of *Statistics of Railways in the United States,* published by the Interstate Commerce Commission. There are somewhat less than 200 roads in Class I (185 were itemized in the 1923 volume). He decided to correlate the total cost of operation with the passenger and freight traffic. However, Class I railways are very different in length, ranging from 21 miles to over 10,000, and this variation must create correlations between the three variables considered which are irrelevant to the main problem. In order to eliminate the disturbing factor, the author used the data given under the heading "averages per mile of road" rather than the totals for each road which are also given. He then correlated these "averages per mile of road" which are the totals of each variable for a given railroad divided by the length of the road. The partial regression coefficients thus computed are expected to measure the average additional cost to the railroads which accompanies a unit increase in the particular service. If a partial regression coefficient is less than the corresponding rate, then the railroads as a whole are adequately paid for their services and make a profit. Otherwise they lose money or, at best, break even. Figures taken from the article are reproduced in Table III.

TABLE III

E = expenses per mile of railroad in $1
F = number of "1000 ton-miles" of freight traffic per mile of railroad
P = number of "1000 passenger-miles" of passenger traffic per mile of railroad

	Average for all roads			Multiple regression equation, expenses being regarded as dependent
	E	F	P	
1919	1865	155.8	20.00	$E = 6.8F + 29P + 221.3$
1921	1936	131.7	16.11	$E = 8.0F + 33P + 353.0$
1922	1858	142.7	15.10	$E = 7.9F + 30P + 279.2$
1923	2022	174.3	15.42	$E = 7.3F + 30P + 285.4$

TABLE IV

L = length of railroad in miles
Z = total expenses in \$10,000
X = total freight traffic in 100,000 ton-miles
Y = total passenger traffic in 100,000 passenger miles

L	Z	X	Y	L	Z	X	Y	L	Z	X	Y
100	61	535	69	100	71	535	69	100	81	535	69
100	61	535	72	100	71	535	72	100	81	535	72
100	61	535	75	100	71	535	75	100	81	535	75
100	61	550	69	100	71	550	69	100	81	550	69
100	61	550	72	100	71	550	72	100	81	550	72
100	61	550	75	100	71	550	75	100	81	550	75
100	61	565	69	100	71	565	69	100	81	565	69
100	61	565	72	100	71	565	72	100	81	565	72
100	61	565	75	100	71	565	75	100	81	565	75
500	90	615	71	500	100	615	71	500	110	615	71
500	90	615	75	500	100	615	75	500	110	615	75
500	90	615	79	500	100	615	79	500	110	615	79
500	90	650	71	500	100	650	71	500	110	650	71
500	90	650	75	500	100	650	75	500	110	650	75
500	90	650	79	500	100	650	79	500	110	650	79
500	90	685	71	500	100	685	71	500	110	685	71
500	90	685	75	500	100	685	75	500	110	685	75
500	90	685	79	500	100	685	79	500	110	685	79
1000	120	627	74	1000	130	627	74	1000	140	627	74
1000	120	627	80	1000	130	627	80	1000	140	627	80
1000	120	627	86	1000	130	627	86	1000	140	627	86
1000	120	700	74	1000	130	700	74	1000	140	700	74
1000	120	700	80	1000	130	700	80	1000	140	700	80
1000	120	700	86	1000	130	700	86	1000	140	700	86
1000	120	773	74	1000	130	773	74	1000	140	773	74
1000	120	773	80	1000	130	773	80	1000	140	773	80
1000	120	773	86	1000	130	773	86	1000	140	773	86
1500	140	700	76	1500	150	700	76	1500	160	700	76
1500	140	700	85	1500	150	700	85	1500	160	700	85
1500	140	700	94	1500	150	700	94	1500	160	700	94

TABLE IV—Continued

L	*Z*	*X*	*Y*	*L*	*Z*	*X*	*Y*	*L*	*Z*	*X*	*Y*
1500	140	800	76	1500	150	800	76	1500	160	800	76
1500	140	800	85	1500	150	800	85	1500	160	800	85
1500	140	800	94	1500	150	800	94	1500	160	800	94
1500	140	900	76	1500	150	900	76	1500	160	900	76
1500	140	900	85	1500	150	900	85	1500	160	900	85
1500	140	900	94	1500	150	900	94	1500	160	900	94
2000	160	800	80	2000	170	800	80	2000	180	800	80
2000	160	800	90	2000	170	800	90	2000	180	800	90
2000	160	800	100	2000	170	800	100	2000	180	800	100
2000	160	900	80	2000	170	900	80	2000	180	900	80
2000	160	900	90	2000	170	900	90	2000	180	900	90
2000	160	900	100	2000	170	900	100	2000	180	900	100
2000	160	1000	80	2000	170	1000	80	2000	180	1000	80
2000	160	1000	90	2000	170	1000	90	2000	180	1000	90
2000	160	1000	100	2000	170	1000	100	2000	180	1000	100
2500	170	900	85	2500	180	900	85	2500	190	900	85
2500	170	900	95	2500	180	900	95	2500	190	900	95
2500	170	900	105	2500	180	900	105	2500	190	900	105
2500	170	1000	85	2500	180	1000	85	2500	190	1000	85
2500	170	1000	95	2500	180	1000	95	2500	190	1000	95
2500	170	1000	105	2500	180	1000	105	2500	190	1000	105
2500	170	1100	85	2500	180	1100	85	2500	190	1100	85
2500	170	1100	95	2500	180	1100	95	2500	190	1100	95
2500	170	1100	105	2500	180	1100	105	2500	190	1100	105
3000	180	1000	90	3000	190	1000	90	3000	200	1000	90
3000	180	1000	100	3000	190	1000	100	3000	200	1000	100
3000	180	1000	110	3000	190	1000	110	3000	200	1000	110
3000	180	1100	90	3000	190	1100	90	3000	200	1100	90
3000	180	1100	100	3000	190	1100	100	3000	200	1100	100
3000	180	1100	110	3000	190	1100	110	3000	200	1100	110
3000	180	1200	90	3000	190	1200	90	3000	200	1200	90
3000	180	1200	100	3000	190	1200	100	3000	200	1200	100
3000	180	1200	110	3000	190	1200	110	3000	200	1200	110

Miss Evelyn Fix was kind enough to prepare Table IV indicating what might have been the raw data regarding the expenditures of the railroads. This table is analogous to Table I. The partial regression coefficients of the same kind as appear in Table III are given in Table V. It will be seen

TABLE V

$z = \frac{Z}{L}$ = expenses per mile of railroad in \$1

$x = \frac{X}{L}$ = number of "1000 ton-miles" of freight traffic per mile of railroad

$y = \frac{Y}{L}$ = number of "1000 passenger-miles" of passenger traffic per mile of railroad

Average for all roads (189):	$\bar{z} = 1943.3$ $\bar{x} = 132.14$ $\bar{y} = 16.043$
Multiple regression equation, expenses being regarded as dependent: $z = 7.994x + 32.87y + 359.7$	

that the conclusions they suggest are similar to those suggested by Table III and entirely contrary to those drawn from the original data of Table IV. In fact, upon inspecting this table it will be seen that, for each fixed size of railroad, the hypothetical expenditures Z are entirely independent of both the total freight traffic X and of the total passenger traffic Y.

The article to which I am referring met with opposition from several authors. However, it is curious that none of the discussants thought that the method of constructing Table III was spurious.

In broad circles of the general public, the opinion still prevails that, in order to conduct statistical studies, one must have enough funds, a few electric calculators and some common sense. Funds and electric calculators are very useful and common sense is just grand. It appears, however, that a little professional education is now and then also useful.

CHAPTER IV

Statistical Estimation

Part 1. Practical Problems and Various Attempts to Formulate Their Mathematical Equivalents

(Based on a conference held in the auditorium of the Department of Agriculture, April 8, 1937, 10 A.M., Mr. Alexander Sturges presiding.)

In this conference I shall try to explain, from the modern point of view, the practical origin of the statistical problem of estimation and some of the early attempts at its solution. The material of the conference falls under four headings. First, under the subtitle "Applicational Roots of the Problem of Estimation," there will be two examples of practical problems. This subsection is followed by two subsections under contrasting subtitles, one on "The Classical Bayes' Approach" and the other on "The Modernized Bayes' Approach." The last subsection is given to the somewhat controversial methods advanced to circumvent the difficulties caused by the absence of exact information regarding the *a priori* distributions of the estimated parameters.

APPLICATIONAL ROOTS OF THE PROBLEM OF ESTIMATION

Practical problems of statistical estimation may be illustrated by the following examples.

Example 1.—We are interested in a certain characteristic ξ of the totality of farms in the United States. This characteristic could be evaluated exactly if we had the necessary data regarding each and every farm. However, the time needed for a one hundred percent survey of farms, and also the cost, would be prohibitive. The best that we can do is this: select a sample of farms for which we will obtain all the pertinent information. Then, the statistical problem of estimation consists in using the data of the sample to evaluate the approximate value of ξ.

Example 2.—As a result of a certain illness, the blood of a patient contains a toxic substance A. The effect of the substance A can be neutralized by giving the patient an injection of a specified chemical B. The treatment will be effective if the dose of B is appropriately adjusted to the average content, say η, of substance A per unit volume of blood of the patient.

The exact value of η could be determined by draining all the blood from the patient and then performing a large scale quantitative analysis of the blood. Since this is impractical, the doctor has to adjust the dosage of his injection, not to the exact value of η, but to the results of analyses of two or three small samples of the blood of the patient. Let $x_1, x_2, \cdots, x_n$ stand for the determinations of A in n samples of blood. The problem of estimation facing the doctor is to use these numbers $x_1, x_2, \cdots, x_n$ in order to obtain a value which, presumably, does not differ very much from η.

Stated in this form, the two problems of estimation just described are not mathematical problems and, therefore, cannot be given a mathematical solution. In fact, it is doubtful whether or not any sort of solution can be offered. Both ξ and η have a strictly defined meaning and can be computed exactly. However, in the practical situation, the data necessary for the evaluation of ξ and η are missing.

In order to arrive at an acceptable solution of the problem of estimation based on calculus of probability, we must begin by translating the problem into the language of probability and by requiring that the method of selecting the sample of farms and the method of determination of the substance A in the samples of blood satisfy certain conditions.

The theory of probability deals with the general question of how frequently this or that event will occur in random experiments of a specified nature. Thus, in order to apply the theory of probability to any domain, this domain must involve some elements of randomness and we must have some information about the nature of the randomness. Thus, if, in the case of the problem regarding the totality of farms in the United States, we are given detailed data for, say, 10,000 farms without any information concerning the method of selection, there is no way in which the theory of probability can be used to estimate ξ. The situation is different if we are told that a sample of 10,000 farms has been drawn at random from the total population of farms in a specified manner. For example, it may be specified that the manner of selecting the sample was such that every possible combination of 10,000 farms had the same chance of being selected. By this, we mean that, if the sampling procedure is repeated many times, then each and every combination of 10,000 farms will be selected approximately with the same frequency. By referring to Part 1, Chapter III, on "Sampling Human Populations," the reader will see that the scheme of sampling just described has been labeled "unrestrictedly random." This is not the only scheme possible and random sampling of farms may be combined with stratification, etc. The essential point is that, in order to apply the theory of probability, the statement of the problem must involve randomness in one form or another. Similarly, in order to use geometry to solve a given practical problem, the conditions of the problem must be

stated in geometrical terms. For example, the question "What is the area of a red triangle?" cannot be solved by plane geometry because of the lack of necessary data expressed in geometrical terms.

Once the random method of sampling farms is specified, the characteristics, say $X_1, X_2, \cdots, X_n$, of the n farms which will appear in the sample become random variables. Ordinarily, the probability distribution of these random variables will depend on the value of ξ and the application of the theory of probability to the problem of estimation becomes possible.

The situation with the second example is quite similar. As long as nothing is known about the determinations of the content of A except that in a given case these determinations gave, say, 3.5 percent, 4.3 percent and 5.1 percent, the theory of probability is helpless to provide anything about the average content of A in the blood of the patient. However, repeated studies of the method of obtaining determinations of A, similar to the work of Matuszewski and Supinska, described in Part 2, Chapter I, may reveal the following.

If the same method is applied many times, then the individual determinations group themselves about the true average content η in a manner characterized by the normal law of frequency. In other words, previous empirical studies may indicate that the relative frequency of determinations falling within any specified interval (a, b) differs but little from the integral

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-(x-\eta)^2/2\sigma^2}\, dx$$

where σ may vary from one patient to another. If so, then the future determinations of A contemplated for a given patient may be considered as random variables following the normal law with unknown mean η and unknown variance σ^2.

Generalizing these remarks relating to two particular examples, we may say that the statistical problem of estimation, to be solvable by means of the theory of probability, must involve the following elements.

(a) There must be one or more random variables, say $X_1, X_2, \cdots, X_n$, particular values of which will be given by future observations. These variables will be described as the observable random variables and, for the sake of brevity, their set will be denoted by a single letter E (the event point).

(b) The probability distribution of the observable random variables must be known to belong to a specified family, say F. Ordinarily, the particular distributions belonging to F are represented by the same formula involving one or more parameters, say $\theta_1, \theta_2, \cdots, \theta_s$, each capable of assuming a certain set of values. Thus, in order to specify completely any one of the

distributions belonging to F, it is sufficient to specify the values of the parameters $\theta_1, \theta_2, \cdots, \theta_s$.

The general problem of statistical estimation consists in devising a method of making assertions regarding the value of one (or more) parameter out of the set $\theta_1, \theta_2, \cdots, \theta_s$, in relation to the particular values of the random variables $X_1, X_2, \cdots, X_n$ which will be furnished by observation.

Example 2 provides a simple illustration of the general situation. Let n be the number of contemplated independent determinations of the content of substance A in samples of blood of a patient. Then the totality E of future observations is represented by n independent random variables, $X_1, X_2, \cdots, X_n$. The empirically supported postulate that each such determination is a normal variable with expectation η and variance σ^2, amounts to postulating that the joint probability density function of E is represented by the formula

$$p_E(x_1, x_2, \cdots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\Sigma(x_i-\eta)^2/2\sigma^2} \tag{1}$$

where η and σ are two parameters with unspecified values and $x_1, x_2, \cdots, x_n$ denote possible values of the random variables $X_1, X_2, \cdots, X_n$. Thus, we may say that, in this particular case, the actual distribution of E is known to belong to the family F of distributions, each characterized by the probability density of the same form (1), with only two parameters, η and σ.

Due to the particular nature of the problem in which η represents the average content of substance A in the blood of the patient, it is possible to assert that η cannot be negative and cannot exceed one. Furthermore, there may be biological reasons insuring that η must lie between even narrower limits. Also, the same kind of argument will be applicable to σ, with the result that it may be taken for granted that its value cannot exceed some specified limits. If we grant the approximation by the normal law, formula (1) and the limits for η and σ summarize our postulated knowledge of the observable random variables $X_1, X_2, \cdots, X_n$. Our interest in the actual value of η leads to the search for a method of using the observed values of $X_1, X_2, \cdots, X_n$ which will be furnished by the chemical analyses to make assertions regarding η.

The words used, to the effect that we search "for a method of making assertions regarding η," do not describe the situation completely. We do not search for just any method of making assertions, but for a method that is, from some convincing point of view, a satisfactory method. Even more, we are likely to prefer the method that is the best of all possible methods.

While there is likely to be general agreement as to the desirability of using the best, or at least a satisfactory, method of making assertions regarding η, there may be difficulty in explaining exactly what properties

a method of estimation should possess in order to qualify as the "best" or as "satisfactory." And without having such an exact explanation, without knowing exactly what we are looking for, it is obviously hopeless to expect that we shall ever find it. If it were possible to devise a method of using the values of the observable random variables to predict exactly and without fail the value of the estimated parameter, then there would be universal agreement that the method in question is the best imaginable. However, it is obvious that, barring some very artificial examples, such a method does not exist and we have to put up with unavoidable errors. For example, whatever the method of using the determinations of the toxic substance A in a few samples of blood, it is obviously impossible to expect that the outcome of estimation will *always* give the true value of η. On the contrary, we may take it for granted that the estimate obtained will always differ from the exact value of η. Similarly, whatever the method of estimating the characteristic ξ of the totality of farms in the United States by the use of a sample, smaller or larger errors of estimation are unavoidable.

This being the case, what should be our definition of a "satisfactory" method of estimation? What should be the definition of the "best" method?

Before attempting to answer these questions, let us consider the possible forms of the assertions regarding the estimated parameter which can be made using the values of the observable random variables. The simplest form is the so-called "point estimate." The method of point estimation of a parameter θ consists of defining a single-valued function, say $\theta^*(E) = \theta^*(X_1, X_2, \cdots, X_n)$, of the observable random variables and, whenever the observations give $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, of making a rule of asserting that $\theta = \theta^*(x_1, x_2, \cdots, x_n)$. The function $\theta^*(E)$ is called the *point estimate* or the *single estimate* of θ.

As already mentioned, in many cases it is more or less hopeless to expect that a point estimate will ever be equal to the true value of θ. In cases of this kind one is naturally interested in the precision of the estimate used. This precision may be usefully characterized by indicating the limits which the error in the estimate, presumably, could not exceed. As a consequence of this tendency, the results of practical investigations are frequently published in the form $\theta^* \pm S$, e.g. 10 ± 1.3, or the like. This form of giving the results of statistical estimation suggests that, while the presumed value of the estimated parameter is 10, there is expected an error of estimation which, however, should not exceed 1.3 either way. It will be noticed that, in effect, this method of estimation amounts to computing from the results of observation not one but two different functions, $\theta^* - S$ and $\theta^* + S$, and asserting that the true value of the parameter θ lies somewhere within the limits from $\theta^* - S$ to $\theta^* + S$.

This procedure is obviously different from that of point estimation. It is described as the estimation by interval. In general terms, the estimation by interval consists of defining not one, but two functions of the observable random variables, say $\underline{\theta}(E)$ and $\bar{\theta}(E)$, and of making a rule of asserting that the true value of θ lies between the limits $\underline{\theta}(x_1, x_2, \cdots, x_n)$ and $\bar{\theta}(x_1, x_2, \cdots, x_n)$, whenever the observations give $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$. The functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ used in this manner are described as the *lower* and the *upper estimate*, respectively. Also, an occasion, it will be convenient to speak of "*theta lower*" and "*theta upper*."

The above two forms of estimation, by single estimate and by interval, are not the only possible methods. In fact, both are particular cases of a more general procedure of estimation by a set. The latter, while theoretically possible, does not seem to have much practical interest. In fact, if it is suggested to use a method of statistical estimation which may lead to the assertion, for example, that the mass of a particle is a number of grams commensurable with $\sqrt{\pi}$ and contained between zero and one, then we may confidently expect that the physicists concerned will show some signs of indignation. For this reason we shall limit our considerations to estimation by a single estimate and by an interval.

With reference to single estimates, roughly one can say that a point estimate, to be satisfactory, should not differ from the estimated quantity "too frequently too much." While intuitive, this statement is obviously too vague to serve as the basis for a theory of estimation. One way of specifying the problem exactly is to reduce it to the problem of estimation by interval. In fact, if this problem is solved satisfactorily, then the point estimate represented by some specified interior point, e.g., by the midpoint of the estimating interval, would probably seem acceptable.

If now we turn our attention to the problem of estimation by interval, we find that this problem is easier to put into exact terms in a manner likely to satisfy the practical statistician. In fact, there is one obvious requirement which any "satisfactory" method of estimation by interval should meet. This is that if it is impossible to arrange that the results of estimation are correct always, we may at least expect them to be correct frequently. In more precise terms, it is natural to require that the estimating interval cover the true value of the estimated parameter with a high relative frequency and that it be possible to fix this frequency in advance. In probabilistic terms this postulate is expressed as follows: (i) *when estimating an unknown parameter θ, the satisfactory lower and upper estimates of θ must have the property that the probability $P\{\underline{\theta}(E) \leqq \theta \leqq \bar{\theta}(E)\}$ be computable and close to unity.* If this probability has a specified large value α, say $\alpha = .99$, then the practical statistician using the estimates $\underline{\theta}(E)$, $\bar{\theta}(E)$ will have the assurance that, in the long run, his assertions regarding the estimated parameters in the form

$\underline{\theta}(E) \leqq \theta \leqq \bar{\theta}(E)$ will be correct about 99 percent of the time, and this is likely to satisfy him. In fact, in this case, each of his assertions regarding the value of θ will be exactly comparable to playing a game of chance with the probability of winning equal to $\alpha = .99$. Naturally, the actual realization of this high frequency of correct assertion regarding θ depends on how closely the postulated properties of the observable random variables agree with the actual conditions of experimentation. Thus, if we postulate that the determinations of the toxic substance A are normally distributed, while in actual fact these determinations have, say, a skew U-shaped distribution, then the interval estimation of η based on the assumption of normality need not give the expected frequency of correct results. However, the basic agreement between the postulates of the theory and the phenomena studied is omnipresent in all problems of application and, in this particular respect, the problems of estimation do not present any sort of exception.

Suppose for a moment that the problem of determining the lower and the upper estimates satisfying requirement (i) is solved and that there is more than one solution. Suppose, for example, that two pairs of functions $\underline{\theta}(E)$, $\bar{\theta}(E)$ and $\underline{\vartheta}(E)$, $\bar{\vartheta}(E)$ both satisfy the condition that

$$P\{\underline{\theta}(E) \leqq \theta \leqq \bar{\theta}(E)\} = P\{\underline{\vartheta}(E) \leqq \theta \leqq \bar{\vartheta}(E)\} = \alpha.$$

Thus, whether the practical statistician uses $\underline{\theta}(E)$ and $\bar{\theta}(E)$ or $\underline{\vartheta}(E)$ and $\bar{\vartheta}(E)$, his assertions regarding the value of the estimated parameter will be correct with exactly the same long run relative frequency α, chosen by himself. In these circumstances, the statistician will be faced with the problem, which we shall denote as problem (ii), of choosing between the estimates $\underline{\theta}(E)$ and $\bar{\theta}(E)$ on the one hand and the estimates $\underline{\vartheta}(E)$ and $\bar{\vartheta}(E)$ on the other. Naturally he will consider the question, which of the two pairs of functions will provide a more exact estimate. If possible, the practical statistician will select for his use the particular pair of functions for which the length of the estimating interval is the least. Should it be impossible to satisfy this condition uniformly so that the selected pair, say $\underline{\theta}(E)$, $\bar{\theta}(E)$, always gives narrower limits for θ than any alternative pair, $\underline{\vartheta}(E)$, $\bar{\vartheta}(E)$, that is,

$$0 \leqq \bar{\theta}(E) - \underline{\theta}(E) \leqq \bar{\vartheta}(E) - \underline{\vartheta}(E),$$

then the practical statistician is likely to formulate some sort of second best requirement substituting "most frequently" for "always" or some such. The essential point in this discussion is that, after finding several estimating intervals capable of covering the true value of the estimated parameter with the same relative frequency, the choice between these estimating intervals will be based on considerations of their length.

You will realize that, compared with the statement of the practical problem of statistical estimation as illustrated in the two examples dis-

cussed at the outset, we have gone a long way toward transforming the practical problem into a mathematical problem. However, even now the problem has not been made entirely precise.

The complete specification of the problem of estimation depends on the assumed conditions. As we have already emphasized, these conditions must imply that the observable random variables $X_1, X_2, \cdots, X_n$ are random variables with a distribution connected in some way with the quantity that one desires to estimate. For example, the observable random variables may be known to be of continuous type and it may be known that their probability density function, say

$$p_{X \mid \theta}(x_1, x_2, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_s),$$

has a known form and depends on some s parameters $\theta_1, \theta_2, \cdots, \theta_s$ the values of which are uncertain. Our problem may be to estimate one (or more) of them, say θ_1. Generally, problems of estimation vary in the amount of knowledge of the distribution of the observable random variables and the connection between the quantity to be estimated and the distribution need not be so simple. However, the assumptions just made are sufficiently illustrative and we shall adhere to them.

In addition to the data regarding the distribution of the observable random variables, the problems of estimation vary in respect to a very important factor which is our assumed knowledge regarding the quantity to be estimated, θ_1, and also regarding such other unknown parameters $\theta_2, \theta_3, \cdots, \theta_s$ as may appear in the probability density $p_{X \mid \theta}(x_1, x_2, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_s)$. In many practical problems, a certain amount of knowledge regarding these parameters is always available. For instance, the conditions of the above example 2 imply that the quantity η is a non-negative number not exceeding unity. Also, there may be some additional items of information which affect the form of the problem of estimation. *The most radical difference in this form depends on whether or not the parameters $\theta_1, \theta_2, \cdots, \theta_s$ are themselves random variables, the distribution of which is known sufficiently to be used in calculations.* In relation to any given problem, this paramount question has to be answered by the practical statistician treating it. Our purpose here will be to describe the nature of the problem of estimation under both sets of conditions, when the unknown parameters are random variables with a known distribution and when they are not.

THE CLASSICAL BAYES' APPROACH

Historically, the first precise treatment of the problem of estimation refers to the case when all the unknown parameters are random variables with a postulated distribution. Therefore, we shall begin our exposition with this particular case. The classical statement and solution of the problem are

based on the famous formula of Bayes.[1] After presenting them, we shall outline a more modern approach to the problem treated under the same conditions.

Consider, then, a set of observable random variables, $X_1, X_2, \cdots, X_n$ with a probability density function $p_{X \mid \theta}(x_1, x_2, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_s)$ which depends on s unknown parameters, $\theta_1, \theta_2, \cdots, \theta_s$. Assume that the conditions of the problem imply that these parameters are also random variables with a probability density function $\Psi_\theta(\vartheta_1, \vartheta_2, \cdots, \vartheta_s)$. Ordinarily, the distribution determined by this function is called the *a priori* distribution of the parameters $\theta_1, \theta_2, \cdots, \theta_s$ and is contrasted with the *a posteriori* distribution obtainable from Bayes' formula, say

$$\Phi(\vartheta_1, \vartheta_2, \cdots, \vartheta_s \mid x_1, x_2, \cdots, x_n) = \frac{\Psi_\theta(\vartheta_1, \vartheta_2, \cdots, \vartheta_s) p_{X \mid \vartheta}(x_1, x_2, \cdots, x_n \mid \vartheta_1, \vartheta_2, \cdots, \vartheta_s)}{\left\{ \int \cdots \int \Psi_\theta(\vartheta_1, \vartheta_2, \cdots, \vartheta_s) \times p_{X \mid \vartheta}(x_1, x_2, \cdots, x_n \mid \vartheta_1, \vartheta_2, \cdots, \vartheta_s)\, d\vartheta_1\, d\vartheta_2 \cdots d\vartheta_s \right\}}. \quad (2)$$

Here the integration in the denominator extends over all systems of values ϑ of the θ's which are compatible with the values $x_1, x_2, \cdots, x_n$ of the observable random variables. Integrating Φ for $\vartheta_2, \vartheta_3, \cdots, \vartheta_s$ over all systems of values compatible with the fixed value ϑ_1 of θ_1, we obtain the *a posteriori* probability density function of θ_1 given the values $x_1, x_2, \cdots, x_n$ of the observable random variables, say

$$\varphi(\vartheta_1 \mid x_1, x_2, \cdots, x_n) = \int \cdots \int \Phi\, d\vartheta_2\, d\vartheta_3 \cdots d\vartheta_s. \quad (3)$$

The product in the numerator of formula (2) represents the joint probability density function of all the parameters $\theta_1, \theta_2, \cdots, \theta_s$ and of all the observable random variables $X_1, X_2, \cdots, X_n$. Integrating it for $\vartheta_1, \vartheta_2, \cdots, \vartheta_s$ for all combinations of their values compatible with the fixed $x_1, x_2, \cdots, x_n$, we obtain the absolute probability density of $X_1, X_2, \cdots, X_n$, say

$$p_X(x_1, x_2, \cdots, x_n) = \int \cdots \int \Psi_\theta(\vartheta_1, \vartheta_2, \cdots, \vartheta_s) p_{X \mid \vartheta}(x_1, x_2, \cdots, x_n \mid \vartheta_1, \vartheta_2, \cdots, \vartheta_s)\, d\vartheta_1\, d\vartheta_2 \cdots d\vartheta_s.$$

This expression appears in the denominator of formula (2).

The function $\varphi(\vartheta_1 \mid x_1, x_2, \cdots, x_n)$ is the basis of the classical procedure of estimating θ_1. Its interpretation is as follows. We visualize a set of cases,

[1] Thomas Bayes: "An essay towards solving a problem in the doctrine of chances." *Phil. Trans., London,* Vol. 53 (1763), pp. 376–398, Vol. 54 (1764), pp. 298–310.

to be described as "human experience" and denoted by H, in which we shall be confronted by the problem of estimating θ_1. In the particular cases which form human experience, the values $\vartheta_1, \vartheta_2, \cdots, \vartheta_s$ of the unknown parameters $\theta_1, \theta_2, \cdots, \theta_s$ vary from case to case and the function $\Psi_\theta(\vartheta_1, \vartheta_2, \cdots, \vartheta_s)$ characterizes the frequency distribution. For example, the relative frequency of cases when θ_1 will fall between any specified limits $a < \theta_1 < b$ is obtainable from Ψ_{θ_1} by integrating it for ϑ_1 between a and b and for $\vartheta_2, \vartheta_3, \cdots, \vartheta_s$ within the extreme limits of their variation.

In parallel with changes in the values of the θ's, the particular cases of human experience will differ in the values, say $x_1, x_2, \cdots, x_n$, assumed by the observable random variables and this variation is characterized by the probability density function $p_{X \mid \vartheta}(x_1, x_2, \cdots, x_n \mid \vartheta_1, \vartheta_2, \cdots, \vartheta_s)$. Now, within human experience H, isolate a part, say $H(x_1, x_2, \cdots, x_n)$, in which the value of X_1 is x_1, the value assumed by X_2 is x_2, etc. Naturally, within the series of cases $H(x_1, x_2, \cdots, x_n)$, the values of the θ's will vary. The above formulae (2) and (3) give the probability density functions relating to the part $H(x_1, x_2, \cdots, x_n)$ of human experience, joint of all the θ's and of θ_1 alone, respectively.

Thus, the exact statement of the classical form of the problem of estimating θ_1 is as follows: *We have observed* $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$*—therefore we appear in part* $H(x_1, x_2, \cdots, x_n)$ *of human experience; what is the most probable value, say* $\hat{\theta}_1(x_1, x_2, \cdots, x_n)$ *of the parameter* θ_1*?* The value $\hat{\theta}_1(x_1, x_2, \cdots, x_n)$ required (called the *a posteriori* most probable value of θ_1 given $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$) is simply that value of ϑ for which $\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)$ is a maximum.

The *a posteriori* most probable values of the estimated parameters have been used extensively as unique estimates since the time of Bayes. Also, once we place ourselves in the specified section $H(x_1, x_2, \cdots, x_n)$ of human experience and limit our consideration to probabilities referring to this section, there is no difficulty in treating the problem of estimation by an interval. In fact, let $\underline{\theta} = \underline{\theta}(x_1, x_2, \cdots, x_n)$ and $\bar{\theta} = \bar{\theta}(x_1, x_2, \cdots, x_n)$ be two numbers which, for a specified α between zero and unity, satisfy the condition

$$P\{\underline{\theta} \leqq \theta_1 \leqq \bar{\theta} \mid x_1, x_2, \cdots, x_n\} = \int_{\underline{\theta}}^{\bar{\theta}} \varphi(\vartheta_1 \mid x_1, x_2, \cdots, x_n) d\vartheta_1 = \alpha. \quad (4)$$

Obviously, there is an infinity of pairs of numbers satisfying this condition and, if $\underline{\theta}$ and $\bar{\theta}$ are to be used to estimate θ_1, it is natural to require that in addition to satisfying (4), the two estimates also minimize the difference

$$\bar{\theta}(x_1, x_2, \cdots, x_n) - \underline{\theta}(x_1, x_2, \cdots, x_n). \quad (5)$$

When the probability density $\varphi(\vartheta_1 \mid x_1, x_2, \cdots, x_n)$ is continuous, the problem of minimizing (5) subject to restriction (4) is trivial and the solution provides the desired estimating interval. This interval will be called *the classical*

Bayes' estimating interval. Its properties are: (a) within the section $H(x_1, x_2, \cdots, x_n)$ of human experience, the frequency of cases where the value of θ_1 will be within the interval $(\underline{\theta}, \bar{\theta})$ equals the number α selected by the statistician himself, and (b) no interval shorter than $(\underline{\theta}, \bar{\theta})$ having the property (a) is in existence. These properties of the classical Bayes' estimating interval may be judged a sufficient justification for using it in practice. However, it is important to be aware of other possibilities which are available when the *a priori* distribution of the unknown parameters is known.

THE MODERNIZED BAYES' APPROACH

In considering these alternative possibilities we should ask ourselves, Why should we refer the probabilities of success in estimation to the section $H(x_1, x_2, \cdots, x_n)$ of human experience? The point of this question is that, even if one is professionally engaged in solving problems of estimation as a matter of daily routine, it will only be most exceptionally that one will be confronted with a set of observations, say $x_1', x_2', \cdots, x_n'$, which has already been observed in the past. Normally, the whole experience of a statistician estimating a given parameter will be composed of cases in which the sets of observations are all different. Thus, this statistician's life experience will consist of cases *each extracted from a different section* $H(x_1, x_2, \cdots, x_n)$ of human experience.

Let us illustrate this by an example. Consider a case in which a contract between a beet sugar factory and a group of beet growers provides for a varying price per ton of beets depending in some way upon the interval used to estimate the average sugar content in a carload of roots. In principle, the sweeter the beets, the higher the price. However, if the estimating interval is broad, a certain decrease in price is allowed due to uncertainty as to the actual sugar content.

In order to determine the price, a sample of beets is drawn out of each carload and several independent determinations, say $X_1, X_2, \cdots, X_n$, of the sugar content are made. These determinations are then used to compute an interval estimating the average sugar content in each carload. Ordinarily, it is assumed that the variables $X_1, X_2, \cdots, X_n$ are independent and follow the normal law of frequency. On this assumption, the probability of observing twice (i.e. for two carloads) the same system of values of the X's is equal to zero. Since the determinations are made only with limited accuracy, strictly speaking, the variables $X_1, X_2, \cdots, X_n$ are not of continuous type and the probability of observing the same system of their values twice (or more) is not exactly equal to zero. Nevertheless, the probability is extremely small and it is safe to say that the experience of the factory will consist of cases where the variables $X_1, X_2, \cdots, X_n$

assume a multitude of different systems of values, with scarcely any repetitions. In these circumstances, it does not seem reasonable to insist that the method of estimating the mean sugar content insures that, *within* each section $H(x_1, x_2, \cdots, x_n)$ of human experience, the probability of covering the true mean sugar content by the estimating interval be exactly equal to the preassigned α. On the contrary, it may be presumed that both the farmers and the administration of the factory will agree that the desirable method of computing the estimating interval should insure (a) that the *overall relative frequency* [contrasted with the relative frequencies relating to each section $H(x_1, x_2, \cdots, x_n)$ separately] of successful estimation be equal to the selected number α close to unity and (b) that, at least on the average, the estimating intervals be as short as possible without infringing condition (a).

It will be seen that here we come to a novel aspect of the problem of estimation. In order to arrive at its exact formulation, it was necessary to realize the fact that, whatever the method used, the outcome of the process of estimation based on some observable random variables has itself the property of being random. It is curious that this fact, noted by Laplace and Gauss, was later forgotten and did not reappear in the literature until in the 1930's.

Upon reflecting on the various practical problems of estimation, it is easy to see that a great many of them resemble the situation implied by the contract between the beet growers and the sugar factory. However, there are examples in which the appropriate point of view on estimation seems to be the classical Bayes' described above. Consider the following situation.

Suppose that the observable random variables $X_1, X_2, \cdots, X_n$ are something like the outcomes of an aptitude test taken by a young man preparing to select a profession for himself. Suppose further that the aptitude test measures exactly the attributes of the individual so that, while $X_1, X_2, \cdots, X_n$ vary from one individual to the next, they are constant for each particular individual. Our final assumption is that the individual's success in the various available professions depends upon the parameter θ_1 to be estimated.

Now consider a particular individual, a Mr. John Frederick Smith, for whom it was found that $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$. It is obvious that Mr. John Frederick Smith's point of view on estimation will be different from that of the administration of the beet sugar factory. The experience of the latter will involve many carloads of beet roots with varying mean sugar content and the important point is to insure overall high frequency

of successes in estimation combined with satisfactory precision. On the other hand, the whole life of Mr. John Frederick Smith will be tied up with just one section of the whole human experience, namely with the section $H(x_1, x_2, \cdots, x_n)$. Therefore, if Mr. John Frederick Smith's actions are to be adjusted at all to outcomes of statistical estimation, it is natural for him to insist on probabilities referring to $H(x_1, x_2, \cdots, x_n)$ rather than to the whole human experience.

To illustrate this point more clearly, let me refer to phenomena of racial discrimination which still infest substantial parts of human society! Imagine that the variables $X_1, X_2, \cdots, X_n$ determine the race of Mr. John Frederick Smith and that he is forced to live in a place where the general circumstances of life of individuals of one race are sharply different from those of another. It is obvious that, having established his racial identity, Mr. John Frederick Smith will be wise to build his own life in conformity with statistical data relating to particular races taken separately, rather than to the overall figures concerned with the total human experience.

The case of John Frederick Smith illustrates, then, the general situation where the classical Bayes' approach to the problem of estimation appears reasonable. The existence of such cases, however, should not blind us, as it did for over a century, to the great mass of other cases in which the restrictiveness of the classical approach can be usefully relaxed. Many important results in this direction, primarily concerned with point estimation, are due to Wald, Wolfowitz, Girshick and others. We will consider the following problem.

Let $X_1, X_2, \cdots, X_n$ denote a set of observable random variables and let θ_1 be the parameter to be estimated. With each system $x_1, x_2, \cdots, x_n$ of possible values of the X's we shall connect a set $\theta(x_1, x_2, \cdots, x_n)$ of possible values of θ_1 to be used for estimating θ_1. Whenever the observations yield $X_1 = x_1', X_2 = x_2', \cdots, X_n = x_n'$, we shall substitute the observed values into the function $\theta(x_1, x_2, \cdots, x_n)$ and assert that the unknown θ_1 is one of the numbers included in $\theta(x_1', x_2', \cdots, x_n')$. Let α be a fixed number, $0 < \alpha < 1$, close to unity.

We shall say that the set $\theta(x_1, x_2, \cdots, x_n)$ is the modernized Bayes' estimating set (MB for short) *corresponding to the confidence coefficient α if it satisfies the following two conditions:*

(1) *The relative frequency of cases within the whole human experience (i.e., the probability) where the set $\theta(X_1, X_2, \cdots, X_n)$ will cover the true value of θ_1 is equal to α;*

(2) *Of all sets satisfying condition* (1) *the set $\theta(X_1, X_2, \cdots, X_n)$ has the smallest expected Lebesgue measure.*

In order to deal with the general case, it is convenient to speak of an estimating set. However, if the reader tries to apply the following theory, he is most likely to find that the modernized Bayes' estimating set reduces to an interval bounded by two functions, say $\theta'(X_1, X_2, \cdots, X_n) \leqq \theta''(X_1, X_2, \cdots, X_n)$.

In order to solve the problem of the MB set, it is sufficient to express the two conditions (1) and (2) in formulae and to apply an easy lemma, occasionally described as the Fundamental Lemma in the theory of optimum tests. Applied to the present case, the lemma asserts the following.

(a) *If $F_1(t_1, t_2, \cdots, t_n)$ and $F_2(t_1, t_2, \cdots, t_n)$ are any two functions integrable over any measurable set of systems of values of the arguments $t_1, t_2, \cdots, t_n$;*

(b) *If w_0 is a set of systems of values of $t_1, t_2, \cdots, t_n$ which contains all systems $(t_1, t_2, \cdots, t_n)$ where*

$$F_1(t_1, t_2, \cdots, t_n) < aF_2(t_1, t_2, \cdots, t_n)$$

and none of those where

$$F_1(t_1, t_2, \cdots, t_n) > aF_2(t_1, t_2, \cdots, t_n);$$

(c) *If w is a measurable set of values of $t_1, t_2, \cdots, t_n$ such that*

$$\int \cdots \int_w F_2 \, dt_1 \, dt_2 \cdots dt_n = \int \cdots \int_{w_0} F_2 \, dt_1 \, dt_2 \cdots dt_n,$$

then

$$\int \cdots \int_{w_0} F_1 \, dt_1 \, dt_2 \cdots dt_n \leqq \int \cdots \int_w F_1 \, dt_1 \, dt_2 \cdots dt_n.$$

In other words, of all sets w for which the integral of F_2 has the same values, the set w_0 ascribes to the integral of F_1 the smallest possible value.

Now let us return to the search for MB sets. For this purpose, consider the systems of possible simultaneous values $(\vartheta, x_1, x_2, \cdots, x_n)$ of the estimated parameter θ_1 and of the observable random variables $X_1, X_2, \cdots, X_n$. In order to visualize these systems, it will be convenient to consider a space S of $n + 1$ dimensions, with axes of coordinates of $x_1, x_2, \cdots, x_n$ and ϑ. The totality of MB sets can be interpreted in the space S as a region (or a set) w_0 containing all points with arbitrary coordinates $x_1, x_2, \cdots, x_n$ and with coordinate ϑ belonging to $\theta(x_1, x_2, \cdots, x_n)$. As usual, W will stand for the whole sample space.

With this interpretation, condition (1) in the definition of MB sets can be expressed by equating to α the integral over w_0 of the joint probability density function of θ_1 and $X_1, X_2, \cdots, X_n$. Thus,

$$\int \cdots \int_W p_X(x_1, x_2, \cdots, x_n) \int_{\theta(x_1,x_2,\cdots,x_n)} \varphi(\vartheta \mid x_1, x_2, \cdots, x_n)\, d\vartheta\, dx_1 \cdots dx_n \quad (6)$$

$$= \int \cdots \int_{w_0} p_X(x_1, x_2, \cdots, x_n)\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)\, d\vartheta\, dx_1 \cdots dx_n$$

$$= \alpha.$$

Similarly, the second condition defining MB sets is expressed by the formula

$$\int \cdots \int_W p_X(x_1, x_2, \cdots, x_n) \int_{\theta(x_1,x_2,\cdots,x_n)} d\vartheta\, dx_1\, dx_2 \cdots dx_n$$

$$= \int \cdots \int_{w_0} p_X(x_1, x_2, \cdots, x_n)\, d\vartheta\, dx_1\, dx_2 \cdots dx_n$$

$$= \text{minimum.}$$

This last formula is due to the fact that the measure of the set $\theta(x_1, x_2, \cdots, x_n)$ is equal to the integral of unity extended over the set. The application of the Fundamental Lemma leads to the conclusion that, in order to determine w_0 it is sufficient to find a constant a and a region w_0 including all points where

$$p_X(x_1, x_2, \cdots, x_n) < a p_X(x_1, x_2, \cdots, x_n)\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)$$

and none of those where

$$p_X(x_1, x_2, \cdots, x_n) > a p_X(x_1, x_2, \cdots, x_n)\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)$$

and such that

$$\int \cdots \int_{w_0} p_X(x_1, x_2, \cdots, x_n)\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)\, d\vartheta\, dx_1 \cdots dx_n = \alpha.$$

Since the probability density p_X is never negative and since we can ignore points where it is zero, it is seen that the region w_0 is defined by the condition

$$\varphi(\vartheta \mid x_1, x_2, \cdots, x_n) \geqq a \quad (7)$$

where a is an appropriate constant. Further on we shall illustrate the procedure in a practical example. It consists in writing down the *a posteriori* probability density function $\varphi(\vartheta \mid x_1, x_2, \cdots, x_n)$ of the estimated parameter and in substituting it into formula (7). This formula must then be solved with respect to ϑ. The solution, in the form of one or more inequalities (combined with equalities) imposed on ϑ, determines the set $\theta(x_1, x_2, \cdots, x_n)$. Obviously, this solution will depend on the chosen a. The value of this constant is adjusted to satisfy condition (6).

In order to illustrate the various concepts discussed we shall now consider some examples. First we shall adopt the classical Bayes' point of view and

illustrate how the *a posteriori* most probable value of a parameter and the classical Bayes' estimating interval depend on the *a priori* distribution of this parameter. Next, on a slightly different example, we will illustrate the relationship between the classical and the modernized Bayes' estimating intervals. In both cases we shall be interested in the conceptual rather than in the practical numerical side of the problem. For this reason, the examples are especially selected so as not to involve technical complications, cumbersome integrals, etc.

We are going to consider n observable random variables $X_1, X_2, \cdots, X_n$, all independent and each known to be uniformly distributed between zero and a positive number θ. We shall assume that our knowledge of this number θ is limited to the double relation, $0 < \theta \leqq 1$, and we shall consider the problem of estimating θ. Thus, in this example, the joint probability density function of all the observable random variables depends on only one unknown parameter, namely θ, and is given by the formula

$$\begin{aligned} p_E(x_1, x_2, \cdots, x_n \mid \theta) &= \frac{1}{\theta^n} \qquad \text{for } 0 \leqq x_1, x_2, \cdots, x_n \leqq \theta \\ &= 0 \qquad \text{elsewhere.} \end{aligned} \tag{8}$$

In order to illustrate the use of Bayes' formula, we shall assume that θ itself is a random variable with the probability density function of the simple form,

$$\begin{aligned} \Psi(\theta) &= m\theta^{m-1} \qquad \text{for } 0 < \theta \leqq 1 \\ &= 0 \qquad \text{elsewhere.} \end{aligned} \tag{9}$$

Here m represents a positive number. Let the letter x without any subscript denote the greatest of the numbers $x_1, x_2, \cdots, x_n$ which may be given by observation as particular values of the variables $X_1, X_2, \cdots, X_n$. The capital letter X without any subscript will denote the random variable defined as the greatest of the $X_1, X_2, \cdots, X_n$. Thus, x is a particular value of X which may be given by observation. The definition of the random variables $X_1, X_2, \cdots, X_n$ implies that $0 \leqq X \leqq \theta$ so that, if the observations have determined a value x of X, then $\theta \geqq x$. Substituting (8) and (9) into (2), we obtain the *a posteriori* probability density of θ, say

$$\begin{aligned} \varphi(\theta \mid x_1, x_2, \cdots, x_n) &= \frac{\theta^{m-n-1}}{\int_x^1 \theta^{m-n-1}\, d\theta} \qquad \text{for } 0 < x \leqq \theta \leqq 1 \\ &= 0 \qquad \text{elsewhere.} \end{aligned} \tag{10}$$

If $m \neq n$, then this formula gives

$$\varphi(\theta \mid x_1, x_2, \cdots, x_n) = \frac{(m-n)\theta^{m-n-1}}{1-x^{m-n}} \qquad \text{for } 0 < x \leqq \theta \leqq 1$$
$$= 0 \qquad \text{elsewhere.} \tag{11}$$

Otherwise, if $m = n$, then

$$\varphi(\theta \mid x_1, x_2, \cdots, x_n) = -\frac{1}{\theta \log x} \qquad \text{for } 0 < x \leqq \theta \leqq 1$$
$$= 0 \qquad \text{elsewhere.} \tag{12}$$

It follows that the *a posteriori* distribution of θ, given $x_1, x_2, \cdots, x_n$, depends effectively only on the greatest x of the $x_1, x_2, \cdots, x_n$. Given the value of x, the most probable value of θ depends on the relation between m and n. If $m = n + 1$, then $\varphi(\theta \mid x_1, x_2, \cdots, x_n)$ is constant within the interval $(x, 1)$,

$$\varphi(\theta \mid x_1, x_2, \cdots, x_n) = \frac{1}{1-x} \qquad \text{for } x \leqq \theta \leqq 1$$
$$= 0 \qquad \text{elsewhere.}$$

Thus, in this particular case, all the numbers of the interval $(x, 1)$ are the *a posteriori* most probable values of θ. Also, in this case, any value may be ascribed to $\underline{\theta}$, subject to the restriction,

$$x \leqq \underline{\theta} \leqq 1 - \alpha(1-x),$$

and then the corresponding value of $\bar{\theta}$ will be

$$\bar{\theta} = \underline{\theta} + \alpha(1-x) \leqq 1.$$

Hence, we have at our disposal an infinity of pairs of estimates varying between the extremes, say

$$\underline{\theta}' = x, \qquad \bar{\theta}' = (1-\alpha)x + \alpha,$$

and

$$\underline{\theta}'' = \alpha x + 1 - \alpha, \qquad \bar{\theta}'' = 1.$$

Any pair of estimates such as these may be used and the consequences will be the same: the probability of the statement regarding the true value of θ which is in the form $\underline{\theta} \leqq \theta \leqq \bar{\theta}$ is equal to the preassigned number α. Also, whatever be the choice of the pair of estimates within the set indicated, the precision of the assertion regarding θ will always be the same because, for all the estimates considered, the difference $\bar{\theta} - \underline{\theta}$ has the same value, namely, $\alpha(1-x)$.

No such arbitrariness of choice exists when $m \neq n + 1$. If $m < n + 1$, then the *a posteriori* probability density of θ decreases as θ varies from x to 1

and the most probable value of θ is x. Also, the left boundary $\underline{\theta}$ of the classical Bayes' estimating interval is $\underline{\theta} = x$. In order to obtain the right boundary $\bar{\theta}$, we have to solve the equation,

$$\int_x^{\bar{\theta}} \varphi(\theta \mid x_1, x_2, \cdots, x_n)\, d\theta = \alpha.$$

If $m \neq n$, then this equation gives

$$\bar{\theta} = [(1 - \alpha)x^{m-n} + \alpha]^{1/(m-n)}.$$

Otherwise, if $m = n$, then $\bar{\theta} = x^{1-\alpha}$.

If $m > n + 1$, then the situation is reversed, the *a posteriori* probability density function of θ increases with the increase of θ from x to 1, and the most probable value of θ is equal to unity, irrespective of the observed value of x. In this case, $\bar{\theta} = 1$ and $\underline{\theta}$ satisfies the equation,

$$\int_{\underline{\theta}}^{1} \varphi(\theta \mid x_1, x_2, \cdots, x_n)\, d\theta = \alpha,$$

which reduces to

$$\underline{\theta} = (\alpha x^{m-n} + 1 - \alpha)^{1/(m-n)}.$$

The purpose of the above discussion is to show that both the single estimate, represented by the *a posteriori* most probable value, and the classical Bayes' estimating interval may depend very strongly on the *a priori* distribution of the estimated parameter θ. In order to emphasize this circumstance, all the results obtained are collected in tabular form.

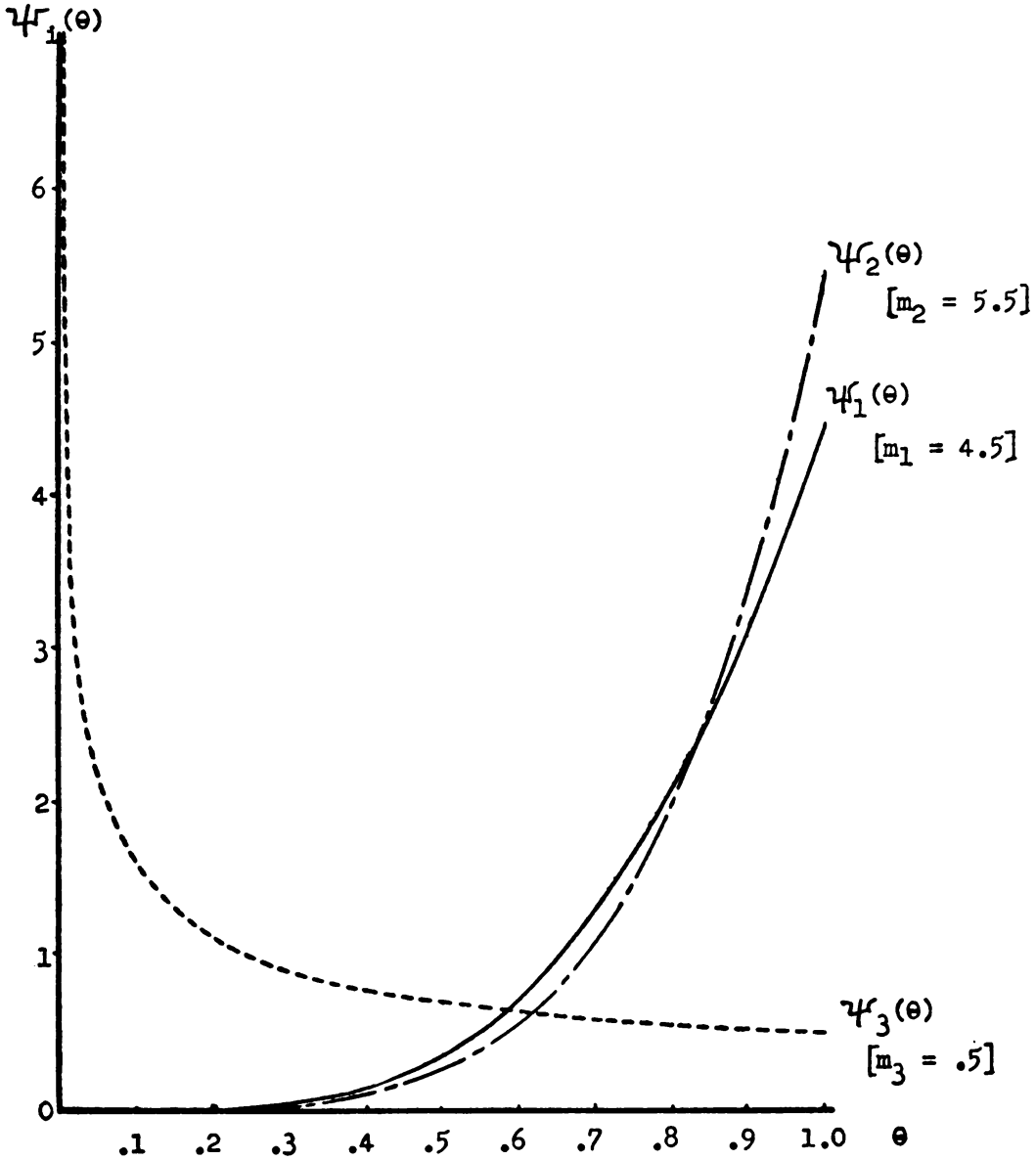
Estimates of θ in relation to m

m	Most probable value $\hat{\theta}(x)$	Theta lower $\underline{\theta}(x)$	Theta upper $\bar{\theta}(x)$
$m < n + 1$			
(a) $m \neq n$	x	x	$[(1 - \alpha)x^{m-n} + \alpha]^{1/(m-n)}$
(b) $m = n$	x	x	$x^{1-\alpha}$
$m = n + 1$	$x \leqq \hat{\theta} \leqq 1$	$x \leqq \underline{\theta} \leqq 1 - \alpha(1 - x)$	$\underline{\theta} + \alpha(1 - x)$
$m > n + 1$	1	$[\alpha x^{m-n} + (1 - \alpha)]^{1/(m-n)}$	1

Figures 1 through 4 illustrate the situation which corresponds to a fixed value of n, $n = 4$, and to three different values of m, $m = 4.5$, 5.5 and .5, respectively. It is seen that for any given x the most probable value of θ and also the classical Bayes' estimating interval depend very much upon

FIGURE 1

A priori distributions of θ

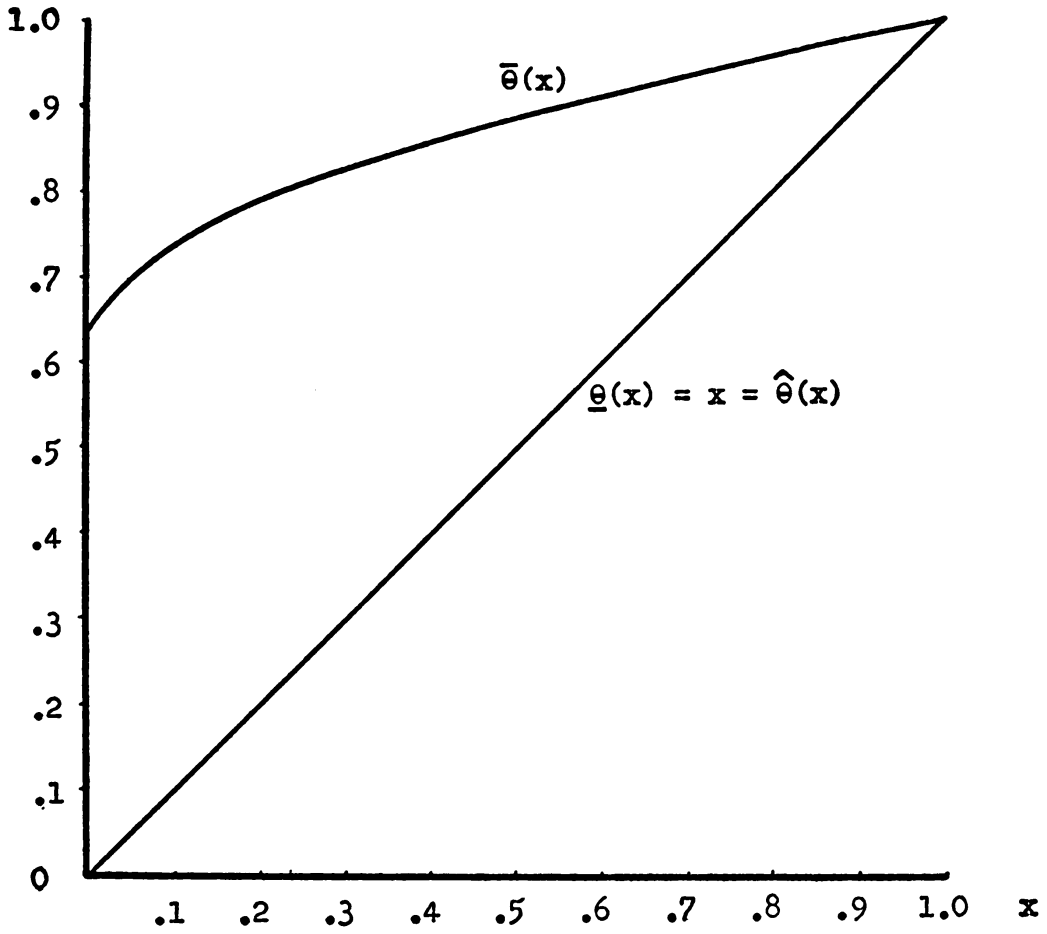


the *a priori* distribution of this parameter. According to the properties of $\Psi(\theta)$, the most probable value of θ may be x itself or unity. The estimating interval may begin at x or end with unity; it may be wide or narrow. The interesting point is that a very substantial change in estimates of θ occurs when the *a priori* distribution $\Psi(\theta)$ changes very moderately, say from $\Psi_1(\theta)$ to $\Psi_2(\theta)$.

FIGURE 2

Classical Bayes' estimating intervals corresponding to $\Psi_1(\theta)$

$m_1 = 4.5; \quad n = 4; \quad a = .8$



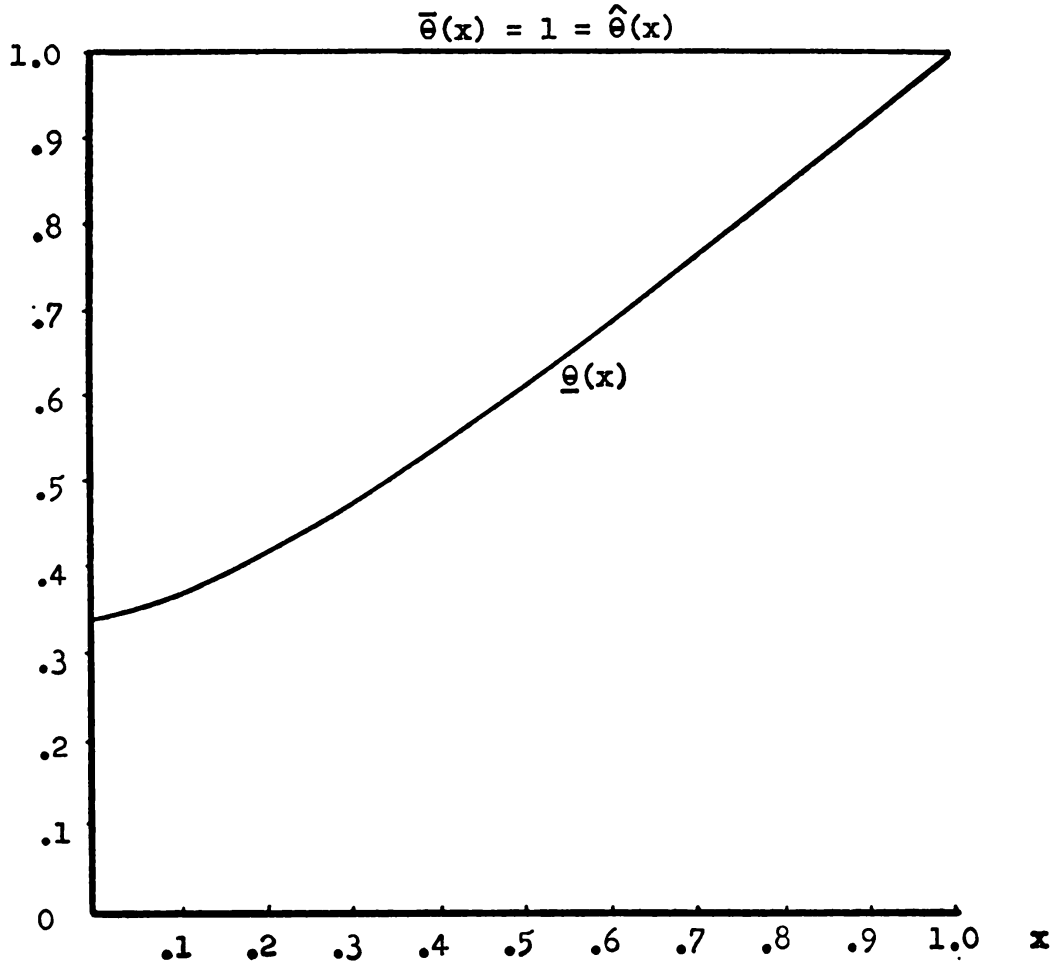
Situations of this kind are quite common and were noticed long ago. When the *a priori* distribution of the estimated parameter is known exactly, there is no difficulty involved. Frequently, however, the *a priori* distribution of the estimated parameter is not known and an effort to use the classical Bayes' approach is combined with the use of a more or less arbitrarily selected function which, it is hoped, approximates the *a priori* probability density. In such cases there may be difficulties.

Now we shall illustrate the relationship between the classical and the modernized Bayes' approach. We shall use the same problem of estimating θ described above, but, in order to simplify the algebra, we shall substitute

FIGURE 3

Classical Bayes' estimating intervals corresponding to $\Psi_2(\theta)$

$m_2 = 5.5; \quad n = 4; \quad \alpha = .8$



$m = 2$ and $n = 3$. Repeating the discussion of the preceding pages, we find easily that the classical Bayes' estimating interval is given by

$$\underline{\theta}(x) = x \qquad \text{and} \qquad \bar{\theta}(x) = \frac{x}{1 - \alpha(1 - x)}, \qquad \text{for } 0 < x \leqq 1. \quad (13)$$

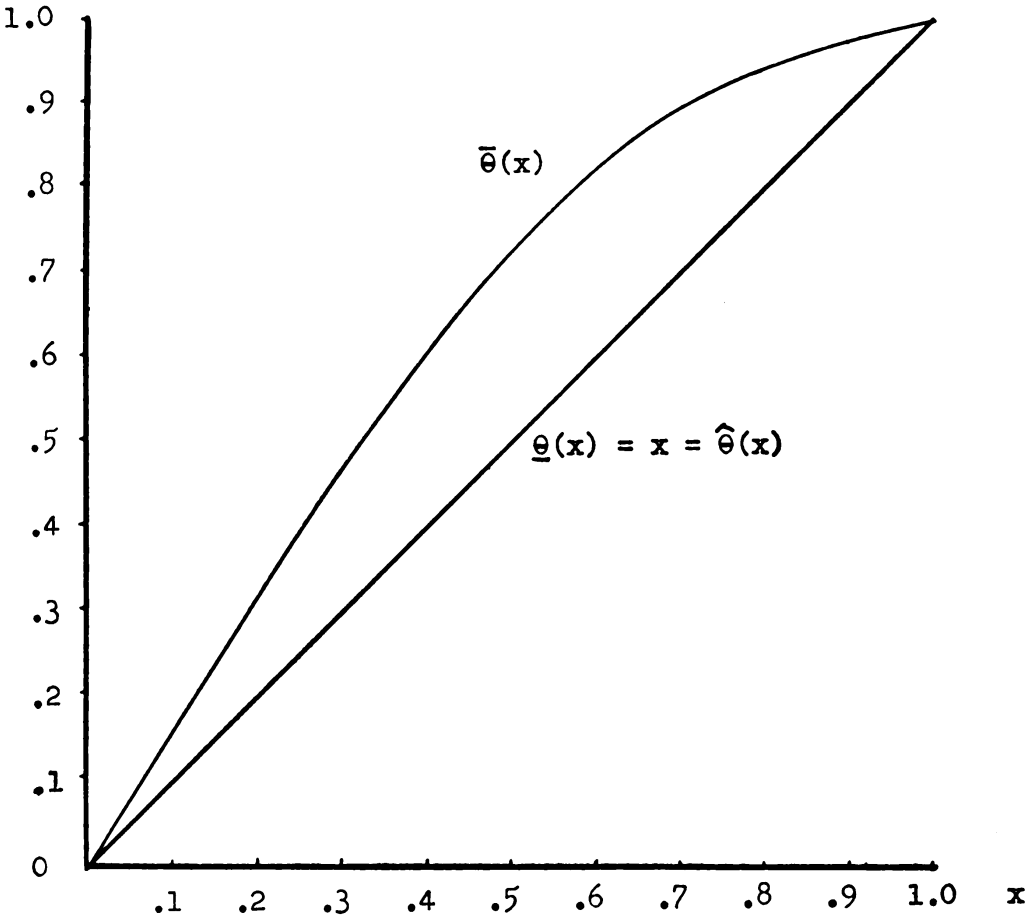
The length of this interval, say $B(x)$, is

$$B(x) = \frac{x}{1 - \alpha(1 - x)} - x, \qquad \text{for } 0 < x \leq 1.$$

FIGURE 4

Classical Bayes' estimating intervals corresponding to $\Psi_3(\theta)$

$m_3 = .5; \quad n = 4; \quad \alpha = .8$



Further, easy calculations indicate that the joint probability density of θ and X is

$$p_{\theta,X}(\vartheta, x) = \frac{6x^2}{\vartheta^2} \qquad \text{for } 0 < x \leqq \vartheta \leqq 1.$$

Integrating this expression for ϑ between limits $x \leqq \vartheta \leqq 1$ we obtain the absolute probability density of X alone,

$$p_X(x) = 6x(1 - x).$$

It follows that the most frequent value of x is one half. Finally, the *a posteriori* probability density of θ is

$$\varphi(\vartheta \mid x) = \frac{p_{\theta,X}(\vartheta, x)}{p_X(x)} = \frac{x}{1 - x}\frac{1}{\vartheta^2} \qquad \text{for } 0 < x < 1 \quad \text{and} \quad x \leqq \vartheta \leqq 1.$$

Turning to formula (7), we see that the MB estimating set is determined by the formula,

$$\frac{x}{1-x}\frac{1}{\vartheta^2} \geqq a,$$

or

$$\vartheta \leqq \sqrt{\frac{1}{a}\frac{x}{1-x}}. \tag{14}$$

We can simplify the further writing somewhat if we substitute $1/t^2$ for a. Then formula (14) will reduce to

$$\vartheta \leqq t\sqrt{\frac{x}{1-x}}. \tag{15}$$

Here a and/or t must be adjusted to the value of α. It will be remembered that conditions of the problem imply that $0 < \vartheta < 1$. Thus inequality (15) implies a real limitation on the value of ϑ only if

$$t\sqrt{\frac{x}{1-x}} \leqq 1$$

or if

$$x \leqq \frac{1}{1+t^2}.$$

Furthermore, (15) is compatible with the necessary condition $x \leqq \vartheta$ only when

$$x \leqq t\sqrt{\frac{x}{1-x}}$$

or

$$x^2 - x + t^2 \geqq 0.$$

This condition will always be satisfied when the roots of the quadratic are complex. This will happen if $t > \frac{1}{2}$. Otherwise (15) will be used only for values of x outside of the interval between the two roots of the quadratic $x^2 - x + t^2$. On the first assumption, namely, that the value of t corresponding to the selected α exceeds one half, the modernized Bayes' estimating interval for θ is, say

$$\begin{aligned} &\theta'(x) = x \leqq \theta \leqq t\sqrt{\frac{x}{1-x}} = \theta''(x) \qquad \text{for } 0 < x \leqq \frac{1}{1+t^2}, \\ &\theta'(x) = x \leqq \theta \leqq 1 = \theta''(x) \qquad \text{for } \frac{1}{1+t^2} \leqq x \leqq 1. \end{aligned} \tag{16}$$

The length of the interval is, say

$$M(x) = \theta''(x) - \theta'(x).$$

Now, let us see how to adjust t to the selected α. For this purpose we write the expression for the probability that the random variables X and θ will satisfy the double relation,

$$P\{\theta'(X) \leqq \theta \leqq \theta''(X)\} = \int_0^1 p_X(x) \int_{\theta'(x)}^{\theta''(x)} p_{\theta \mid x}(\vartheta)\, d\vartheta\, dx$$

$$= \int_0^{1/(1+t^2)} p_X(x) \int_x^{t[x/(1-x)]^{1/2}} p_{\theta \mid x}(\vartheta)\, d\vartheta\, dx$$

$$+ \int_{1/(1+t^2)}^1 p_X(x) \int_x^1 p_{\theta \mid x}(\vartheta)\, d\vartheta\, dx.$$

It will be seen that this probability is an increasing function of t. Upon substituting the expressions of $p_X(x)$ and of $p_{\theta \mid x}(\vartheta)$ given above and upon performing the integration, we find that

$$P\{\theta'(X) \leqq \theta \leqq \theta''(X)\} = \frac{4t^4 + 11t^2 + 9}{4(1+t^2)^2} - \frac{3}{8t}\left(\arcsin \frac{1-t^2}{1+t^2} + \frac{\pi}{2}\right)$$

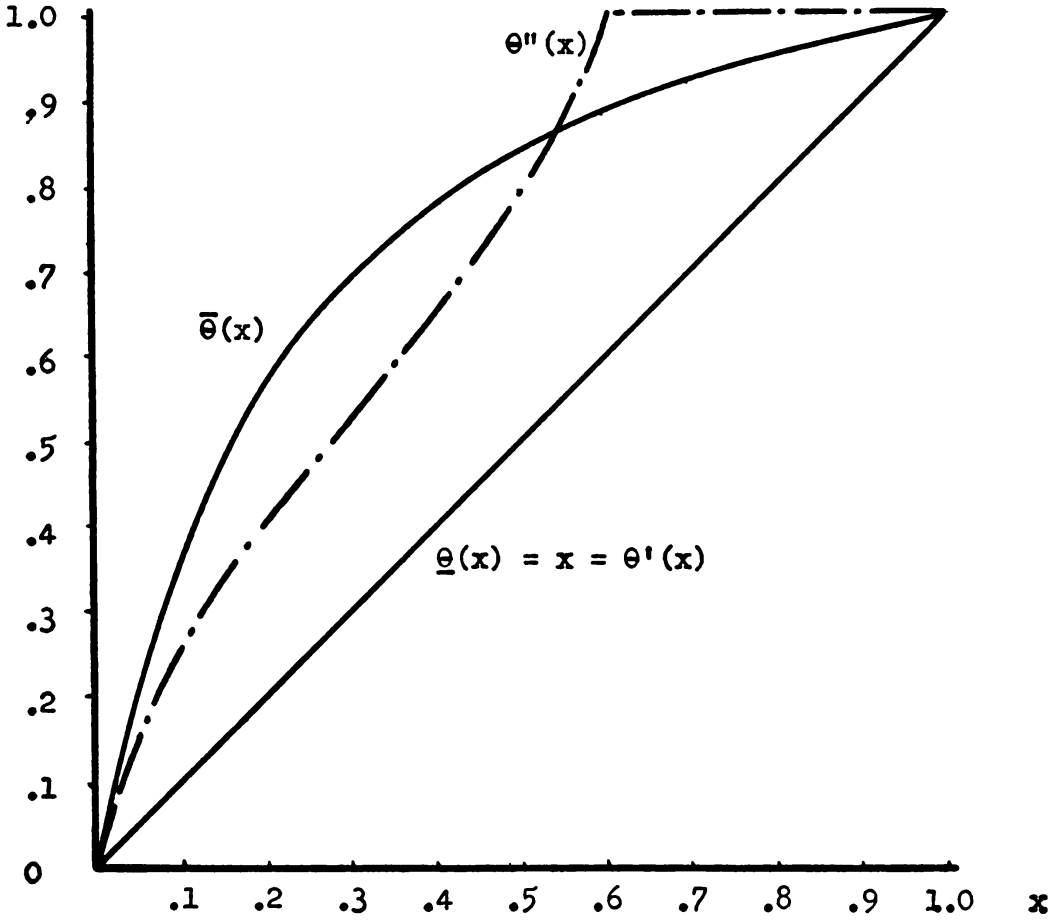
$$\text{for } \tfrac{1}{2} \leqq t. \quad (17)$$

The requisite value of t can be found by equating this expression to α and by solving with respect to t. First, however, we must assure ourselves that this value exceeds one half. For this purpose we first decide on $\alpha = 0.8$. Upon substituting $t = \frac{1}{2}$ into (17), we obtain the value $P = .2593$. Hence the requisite value of t must be greater and formula (17) may be used to establish it. Easy interpolation gives $t = .79614$. This completes the determination of the modernized Bayes' estimating interval for θ. Figure 5 presents both the classical Bayes' estimating interval corresponding to formula (13) with $\alpha = 0.8$ and the modernized, corresponding to (16) with $t = .79614$. If he uses either, the statistician may be sure that he will be correct in about 80 percent of the cases. Using the classical solution (13) he may also be certain that, should it be possible to isolate enough cases of the general human experience in which X has the same value x, the relative frequency of successes in this section $H(x)$ of human experience would also be equal to α. In addition, no shorter interval having this property can be found. If he uses the modernized solution, no general statement regarding any particular section $H(x)$ can be made. The inspection of the graphs on Figure 5 indicates that the MB intervals are sometimes shorter and sometimes longer than the classical ones. In using the MB

FIGURE 5

Comparison of the classical and the modernized Bayes' estimating intervals

$m = 2; \quad n = 3; \quad \alpha = .8$



intervals, the statistician may be certain that within the whole human experience he will also be right in about 80 percent of all cases. Furthermore he may be certain that no other estimating intervals corresponding to the same α will have their long range average length shorter than the intervals computed from (16). Whatever the interval, say $\theta_1(x) \leqq \theta_2(x)$, the expectation of its length is computed from the formula,

$$E[\theta_2(x) - \theta_1(x)] = \int_0^1 [\theta_2(x) - \theta_1(x)]p_X(x)\,dx.$$

Upon substituting into this formula alternatively the expressions (13) and (16), it is found

$$E[B(X)] = 1 - \frac{\alpha}{2} - (1-\alpha)^{3/4}\Big|_{\alpha=.8} = .2868,$$

$$E[M(X)] = \frac{3}{8}t\left[\arcsin\frac{1-t^2}{1+t^2} + \frac{\pi}{2}\right] - \frac{2+t^2+t^4}{4(1+t^2)^2}\Big|_{\alpha=.8} = .2522.$$

It is seen that, by sacrificing the requirement underlying the classical approach that the probability of success in estimation be equal to $\alpha = 0.8$ separately within each section $H(x)$ of human experience, and by requiring that this probability equal α within the whole human experience, the average length of the estimating interval can be reduced roughly by 12 percent of its original value. Naturally, this particular percentage is characteristic of the particular problem considered. The important point to remember is that, outside of the category of cases exemplified by the above situation of Mr. John Frederick Smith, the modernized Bayes' estimating intervals will ordinarily be, on the average, shorter than the classical ones corresponding to the same frequency of successful estimation.

As we have already mentioned, the problem of the modernized Bayes' estimating intervals is akin to the theory treated by Wald, Wolfowitz, Girshick and others. However, the theory that is treated by these authors is much deeper and refers to the case in which the *a priori* distribution of the estimated parameter is uncertain. Also this theory is mainly concerned with point estimation.[2]

DIFFICULTIES CAUSED BY UNCERTAINTY REGARDING THE A PRIORI DISTRIBUTION AND ATTEMPTS TO CIRCUMVENT THEM

As we have already mentioned, all the above discussion applies to cases where the *a priori* distribution of the estimated parameter is exactly known. This distribution must be implied by the conditions of the particular problem under consideration. Cases of this kind exist, particularly in genetics where the postulate of the Mendelian Law implies everything, the randomness of the observable variables, the class to which their distribution belongs, the randomness of the estimated parameters and their *a priori* distribution.

Unfortunately, situations of this nature are extremely rare and, in problems of a more common type, various difficulties arise. The randomness of the estimated parameter requires a postulate entirely independent of the one which concerns the observable random variables. Moreover, on occasion one feels reluctant to admit that the parameters are random variables. Finally, even if the randomness of the parameters is postulated,

[2] For illustrations of the problems currently treated, see the article by J. L. Hodges, Jr., and E. L. Lehmann: "Some problems in minimax point estimation," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 182–197.

their distribution is unknown so that the conditions of the practical problem considered do not include information about the nature of the function $\Psi(\theta_1, \theta_2, \cdots, \theta_s)$ which plays an important role in formula (2). Strictly speaking, then, in cases of the kind described, the formulae discussed above, giving the most probable value of the parameter and the classical or the modernized Bayes' estimating interval, are not applicable because of lack of necessary data.

This embarrassing circumstance was noticed quite some time ago and there have been various attempts made to overcome the difficulty. Most of these attempts have the same theoretical weakness: they are not solutions of a mathematical problem using the data which are directly implied by the practical problem considered; instead, they are excuses or alibis for applying the attractive formula (2) even though the conditions of the problem studied do not provide the necessary data to substitute in formula (2). Naturally, these theoretical weaknesses are accompanied by corresponding practical defects.

The first attempt to obviate the difficulty caused by the lack of information regarding the probabilities *a priori* consisted in the formulation of the so-called "principle of insufficient reason." Roughly, this principle asserts that, whenever there is no good reason to believe that some particular possible values of the estimated parameter are more probable than others, then it is legitimate to substitute in formula (2)

$$\Psi(\theta_1, \theta_2, \cdots, \theta_s) = C = \text{constant}.$$

There are no laws, as yet, prohibiting the calculation of any formulae, and I would be the last to suggest that such laws should be introduced. Thus, I have not the slightest intention of questioning the legitimacy of the substitution suggested. On the other hand, I wish to point out that, in cases where the conditions of the actual problem do not imply $\Psi(\theta_1, \theta_2, \cdots, \theta_s) = C$, and where the substitution of C instead of $\Psi(\theta_1, \theta_2, \cdots, \theta_s)$ is made on the basis of the principle alone, the results of further calculations using formula (2) need not have the clear frequency interpretation discussed above. In particular, the most probable value of the parameter computed using the principle of insufficient reason need not coincide with the value of θ which is most frequent in the sequence of cases $H(x)$. Furthermore, the estimating interval computed using the principle of insufficient reason need not contain the true value of θ in the stated proportion α of the sequence of cases $H(x)$.

This point is well illustrated in the example described above. Following the principle of insufficient reason, we should put $m = 1$ which, with $n = 4$, would lead to the conclusion that the *a posteriori* most probable value of θ is x, the greatest of the four values of the observable random variables given by observation. However, if it happens that the true *a priori* distri-

bution of θ is the function $m\theta^{m-1}$ with $m > 5$, then, irrespective of the observed value of x, the value most frequently assumed by θ in any sequence $H(x)$ will be unity. Furthermore, the presumed "most probable" value equal to x will be the least frequent value of θ. A similar disappointment would result from the application of the Bayes' estimating interval based on the principle of insufficient reason.

In addition to the above disadvantages, the principle of insufficient reason is difficult to apply when the set of the possible values of the estimated parameter is unbounded, for example when the parameter θ is capable of assuming any positive value $0 < \theta < \infty$, or any real value $-\infty < \theta < +\infty$. In cases of this kind the probability density function cannot be represented by a constant because of the restriction that the integral of the probability density function extended from $-\infty$ to $+\infty$ must be equal to unity. Strange as it may seem, some of the protagonists of the subjective theory of probability who adhere to the principle of insufficient reason, are not disturbed by this fact.

In this connection a modification of the principle of insufficient reason should be mentioned. According to the new principle, formula (2) may be legitimately computed by substituting for the unknown $\Psi(\theta_1, \theta_2, \cdots, \theta_s)$ some function, not necessarily a constant, but a special function invented for this particular purpose and representative of "the state of mind" of the statistician who lacks any knowledge of what the values of the parameters $\theta_1, \theta_2, \cdots, \theta_s$ might be.

Thus, for example, when one deals with normally distributed variables having an unknown variance σ^2, in the absence of any definite information as to what the *a priori* distribution might be, it is suggested that one use the formula, say

$$\Psi_J(\sigma) = \frac{c}{\sigma},$$

where c is a constant.

The reason for suggesting this particular function seems to be the following. Let t be a positive number. If we try to answer the question of the relation between the probability of $\sigma < t$ and the probability of $\sigma > t$, the suggested form of $\Psi_J(\sigma)$ has the advantage of not providing any answer. In fact, treating $\Psi_J(\sigma)$ as the probability density of σ over the whole range of possible values of σ, from zero to $+\infty$, we may attempt to compute the desired probabilities by taking the integrals

$$P\{0 < \sigma < t\} = c\int_0^t \frac{d\sigma}{\sigma} = c \log \sigma \Big|_0^t,$$

$$P\{t < \sigma < +\infty\} = c\int_t^\infty \frac{d\sigma}{\sigma} = c \log \sigma \Big|_t^\infty.$$

It happens, however, that both of these integrals diverge and hence that there is no real number representing either of them. Thus, it is impossible to answer the question whether it is more probable that $\sigma < t$ or that $\sigma > t$. Allegedly, this corresponds exactly to our state of mind regarding σ, namely, to the complete lack of knowledge regarding its value. From this point of view, one might regret perhaps that the integral of $\Psi_J(\sigma)$ taken between any positive limits, $0 < a < b$, converges so that, for example, it appears possible to compare the probabilities $P\{.1 < \sigma < .2\}$ and $P\{1 < \sigma < 2\}$ and to find them equal. This circumstance does not seem to be consistent with the complete ignorance of the value of σ which was postulated.

A much pleasanter attempt to deal with the lack of probabilities *a priori* consists in an effort to estimate them empirically. This method was used by many authors but recently it was explicitly advocated by R. v. Mises.[3] We may illustrate its use and also its shortcomings on the two examples mentioned at the beginning of this conference. Thus the doctor who specializes in treating patients with an excessive content of chemical A in their blood may keep records of his determinations of chemical A. According to some method, perhaps similar to the one used by Mrs. Tang in estimating the distribution of the true sugar excess in varieties of sugar beet, the doctor may establish a function, say $\Psi_M(\eta)$, which represents approximately the true distribution of η in the population of persons ill with the particular disease, of whom his office patients, in the course of the last year or so, were a sample. He may then use this function for purposes of estimating η during the following year.

Undoubtedly, this method of approach is far more realistic than the invention of *a priori* distributions without any recourse to actual phenomena. If it happens both that the population of ill persons does not change from one year to another and that the growing reputation of the doctor does not produce a change in the recruitment of his patients, then the function representing the probability density of η in one year will be valid for the next, and the doctor's adjustment of the dose of the drug B will not be more inaccurate than expected. However, it is common knowledge that the conditions of health change from one year to another and from one vicinity to the next and it is these changes that are the danger points of the doctor's proposed procedure. In fact, conditions of health may be presumed just as variable in time as the conditions of breeding studied by Mrs. Tang in 1937. Also, there is another danger, connected with the unavoidable inaccuracies in estimating the *a priori* distribution of η using past experience. To illustrate this point, I call your attention to Figure 1 and to Figures 2 and 3. It is not impossible that the unavoidable random

[3] Richard von Mises: "On the correct use of Bayes' formula." *Annals of Math. Stat.*, Vol. 13 (1942), pp. 156–165.

errors involved in using past experience may result in assuming that the *a priori* probability density function of θ is represented by $\Psi_1(\theta)$ whereas, in actual fact, the true probability density function of this parameter is $\Psi_2(\theta)$. Figures 2 and 3 show, then, that the *a posteriori* conclusions based on $\Psi_1(\theta)$ will be very different from the realities implied by $\Psi_2(\theta)$.

While the suggested procedure of estimating the *a priori* distribution using past experience involves some dangers, still it is applicable in all cases where the same problem of estimation appears again and again and provides the opportunity of collecting a reasonable amount of data. The first example used in this conference illustrates a category of problems in which the procedure is not applicable. In fact, while in recent decades, the governments of all civilized countries have made repeated attempts to study various phases of their economy, including farming, the number of observations made in the past is plainly insufficient to provide any sort of approximation to the *a priori* distribution of a characteristic like the hypothetical characteristic ξ of the totality of farms. In addition, it is well known that the economic processes are rather rapid and the totality of farms in 1950 is a population entirely different from that in 1940 or in 1930. These populations and their characteristics are external marks of the current economic development with its periods of booms and recessions, and this is just the reason why, short of a comprehensive probabilistic theory of national economy, I personally am reluctant to consider ξ as a random variable. The postulation of a definite probability distribution of ξ would seem to be even less appropriate.

Studies of populations are usually made on relatively large samples, certainly in hundreds and frequently in thousands or in tens of thousands. Cases of this kind have inspired two great mathematicians, S. Bernstein and, apparently somewhat later, R. v. Mises, to prove a very interesting theorem regarding the properties of the *a posteriori* distribution when the number of independent observations is indefinitely increased.

Bernstein obtained his result some time before 1915. In fact, in 1915 he described it in his lectures on probability which I had the good fortune to attend. R. v. Mises published his result in 1919 [4] in *Mathematische Zeitschrift*. It must have been proved a few years earlier, thus, at about the same time as S. Bernstein's result. Both results are to the general effect that, when the *a priori* distribution of a given parameter and also the distribution of the observable random variables satisfy certain conditions of regularity, then the standardized *a posteriori* distribution of the estimated parameter, given n independent observations, tends, as $n \to \infty$, to the normal distribution with zero expectation and unit variance. Thus, no matter

[4] Richard von Mises: "Fundamentalsätze der Wahrscheinlichkeitsrechnung." *Math. Zeit.*, Vol. 4 (1919), pp. 1–97.

which particular function Ψ satisfying the conditions of regularity we substitute into formula (2), if n is sufficiently large, the results of computing formula (2) will be approximately the same. Namely, whatever $a < b$, the *a posteriori* probability that the difference between the true value of θ and its *a posteriori* expectation will be between $a\sigma_\theta$ and $b\sigma_\theta$ will be approximately equal to

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx,$$

where σ_θ^2 denotes the *a posteriori* variance of θ.

It should be mentioned that the original paper of R. v. Mises did not enumerate the restrictions needed for the above conclusion. This gap was later filled by J. Hosiasson.[5]

While the Bernstein-v. Mises theorem is a very interesting result, revealing important properties of the *a posteriori* distributions, it has definite shortcomings if it is treated as the basis for extensive applications of the Bayes' formula. First of all, many problems of estimation arise in which the number n of observable random variables is small and it is more or less hopeless to rely on an asymptotic result based on passage to the limit with $n \to \infty$. Second, the Bernstein-v. Mises theorem requires that certain regularity conditions be satisfied, and it happens that these conditions are far from being met universally. For example, they are not satisfied in the example of n independent variables following distribution (8). As a result, the *a posteriori* distribution of θ is positive only on the interval $(x,1)$ and, for large values of n, is monotonically decreasing as θ varies from x to 1. Obviously, it cannot be made to approach a normal limiting distribution by a mere process of standardization.

The third shortcoming is somewhat delicate. In order to explain it I call your attention to the description of the theorem given above: "no matter which particular function Ψ satisfying the conditions of regularity we substitute in formula (2), if n is sufficiently large, then . . ." [This statement is made on the premise that the appropriate conditions of regularity are also met by $p_E(x_1, x_2, \cdots, x_n \mid \theta_1, \cdots, \theta_s)$.] Thus, the theorem asserts that for every function Ψ of the specified category there exists appropriately large values of n with which the true *a posteriori* distribution (standardized) differs but little from the normal law.

However, the theorem does not assert that sufficiently large values of n can be found such that, whatever *a priori* distribution of the specified broad category we take, the difference between the true *a posteriori* distribution

[5] Janina Hosiasson: "Quelques remarques sur la dépendance des probabilités *a posteriori* de celles *a priori*." *Comptes-rendus, Premier Congrès des Math. des Pays Slaves, Warszawa, 1929* (1930), pp. 375–382.

and the normal limit can be neglected. The Bernstein-v. Mises theorem does not assert this and the assertion is not true. It follows that, in spite of all its interest, this theorem cannot be considered as a universal justification for the use of Bayes' formula in all cases where the *a priori* distribution of a parameter is uncertain.

In the history of methods of estimation, the two principles of best unbiased estimates and of maximum likelihood estimates play a role apart. Both were used by Gauss and by many authors thereafter. The principle of best unbiased estimates was never formulated unambiguously as a principle but simply came into frequent use, partly because it is easily applied in a broad category of cases and partly because it has important advantages as proved by Gauss and later popularized by Markoff.

The principle of maximum likelihood was definitely proclaimed as a principle. This was done by R. A. Fisher in a number of his writings from which I shall give you a few quotations. However, probably feeling the weakness of a dogma, Fisher [6] tried to support the dogma by rational arguments. In this he was very successful and guessed a number of important properties of the maximum likelihood estimates. Under suitable restrictions these properties were subsequently proved, with increasing rigor and generality, by Harold Hotelling,[7] J. Doob [8] and, finally, A. Wald.[9]

Thus, the maximum likelihood estimates will be considered from two different points of view. First, we shall consider them from the point of view in which the *principle* of maximum likelihood is understood to be a command to use these particular estimates for the sole reason that they maximize the likelihood function. The second time we shall consider the use of the same estimates prompted not by the principle but by an understanding of the properties which they possess in certain specified cases. In this, the maximum likelihood estimates will appear to play a role somewhat similar to that of best unbiased estimates, depending upon the prior solution of the problem of estimation by interval, i.e. upon the solution which is independent of any assumption regarding the probabilities *a priori*.

Here is a word of warning. The justification for the use of best unbiased and maximum likelihood estimates just mentioned is merely *a justification*. It is not intended to suggest that this is the only justification possible. We

[6] R. A. Fisher: "On the mathematical foundations of theoretical statistics." *Phil. Trans. Roy. Soc., London,* Ser. A, Vol. 222 (1922), pp. 309–368.

[7] Harold Hotelling: "The consistency and ultimate distribution of optimum statistics." *Trans. Am. Math. Soc.,* Vol. 32 (1930), pp. 847–859.

[8] Joseph Doob: "Probability and statistics." *Trans. Am. Math. Soc.,* Vol. 36 (1934), pp. 759–775.

[9] A. Wald: "Note on the consistency of the maximum likelihood estimate." *Annals of Math. Stat.,* Vol. 20 (1949), pp. 595–601.

shall discuss this in the next conference after presenting the non-Bayes' solution of the problem of estimation by interval.

Fisher's dogmatic attitude towards maximum likelihood estimates may be illustrated by the following quotations, in which the more relevant passages are italicized.

> The rejection of the theory of inverse probability [of the use of Bayes' formula with an invented *a priori* distribution: J. N.] was for a time wrongly taken to imply that we cannot draw, from knowledge of a sample, inferences respecting the corresponding population. Such a view would entirely deny validity to all experimental science. What has now appeared is that *the mathematical concept of probability is,* in most cases, *inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term* **"Likelihood"** *to designate this quantity; . . . .* (R. A. Fisher: *Statistical Methods for Research Workers.* 11th ed. Oliver and Boyd, London, 1950, p. 10.)
>
> The fact that the concept of probability is adequate for the specification of the nature and extent of uncertainty in these deductive arguments is no guarantee of its adequacy for reasoning of a genuinely inductive kind. . . . More generally, however, a mathematical quantity of a different kind, which I have termed *mathematical likelihood,* appears to take its place as a measure of rational belief when we are reasoning from the sample to the population. (R. A. Fisher: "The logic of inductive reasoning." *Jr. Roy. Stat. Soc.,* Vol. 98 (1935), p. 40.)

These quotations illustrate a difference between Fisher's attitude towards probability and my own. For Fisher, probability appears as a measure of uncertainty applicable in certain cases but, regretfully, not in all cases. For me, it is solely the answer to the question, "how frequently this or that happens."

Now, here are a few quotations from Fisher illustrating his non-dogmatic attitude toward the principle of likelihood.

> Obviously the claim that the likelihood possesses these properties, and provides a rational basis for exact inference, can only be made in the light of a theory of estimation applicable to finite samples. In (2) [10] I have developed such a theory, and have demonstrated that the most likely value of x, that is, the particular estimate found by the method of maximum likelihood, possesses uniquely those sampling properties which are required of a satisfactory estimate. (R. A. Fisher: "Inverse probability and the use of Likelihood." *Proc. Cambridge Phil. Soc.,* Vol. 28 (1932), pp. 257–261.)

Here, then, there is no contention that the likelihood function is in itself a measure of confidence in a given value of a parameter. On the other hand, it is claimed that it is advantageous to use the maximum likelihood estimates *because* they have some desirable properties. Some of the desirable and undesirable properties of an estimate are described as follows.

[10] R. A. Fisher: "On the mathematical foundations of theoretical statistics." *Phil. Trans. Roy. Soc., London,* Ser. A, Vol. 222 (1922), pp. 309–368.

If we calculate a statistic, such, for example, as the mean, from a very large sample, we are accustomed to ascribe to it great accuracy; and indeed it will usually, but not always, be true, that if a number of such statistics can be obtained and compared, the discrepancies among them will grow less and less, as the samples from which they are drawn are made larger and larger. In fact, as the samples are made larger without limit, the statistic will usually tend to some fixed value characteristic of the population, and, therefore, expressible in terms of the parameters of the population. If, therefore, such a statistic is to be used to estimate these parameters, there is only one parametric function to which it can properly be equated. If it be equated to some other parametric function, we shall be using a statistic which even from an infinite sample does not give the correct value; it tends indeed to a fixed value, but to a value which is erroneous from the point of view with which it was used. Such statistics are termed **Inconsistent** Statistics; except when the error is extremely minute, as in the use of Sheppard's adjustments, inconsistent statistics should be regarded as outside the pale of decent usage. (R. A. Fisher: *Statistical Methods for Research Workers.* 11th ed. Oliver and Boyd, London, 1950, p. 11.)

With this preference of Fisher not to use inconsistent statistics, I perfectly agree. When one intends to estimate a parameter θ, it is definitely not profitable to use an inconsistent estimate.

Consistent statistics, on the other hand, all tend more and more nearly to give the correct values, as the sample is more and more increased; at any rate, if they tend to any fixed value it is not to an incorrect one. In the simplest cases, with which we shall be concerned, they not only tend to give the correct value, but the errors, for samples of a given size, tend to be distributed in a well-known distribution . . . known as the Normal Law of Frequency of Error, or more simply as the **normal distribution.** The liability to error may, in such cases, be expressed by calculating the mean value of the squares of these errors, a value which is known as the **variance;** and in the class of cases with which we are concerned, the variance falls off with increasing samples, in inverse proportion to the number in the sample.

Now, for the purpose of estimating any parameter, such as the centre of a normal distribution, it is usually possible to invent any number of statistics such as the arithmetic mean, or the median, etc., which shall be consistent in the sense defined above, and each of which has in large samples a variance falling off inversely with the size of the sample. But for large samples of a fixed size the variance of these different statistics will generally be different. Consequently, a special importance belongs to a smaller group of statistics, the error distributions of which tend to the normal distribution, as the sample is increased, with the least possible variance. We may thus separate off from the general body of consistent statistics a group of especial value, and these are known as **efficient** statistics.

. . . . .

The researches of the author have led him to the conclusion that an efficient statistic can in all cases be found by the **Method of Maximum Likelihood;** that is, by choosing statistics so that the estimated population should be that for which the likelihood is greatest. (R. A. Fisher: *Statistical Methods for Research Workers.* 11th ed. Oliver and Boyd, London, 1950, pp. 11–14).

Here, again, I agree unreservedly with Fisher that, when several consistent estimates of the same parameter are available, all tending to be

normally distributed, the one with the smallest variance is preferable to others. Consequently, whenever the method of maximum likelihood yields estimates which are both consistent and efficient, this circumstance (but not the principle) may be considered an inducement to use the maximum likelihood estimates. On the other hand, if and when the maximum likelihood estimates are either inefficient or are outside "the pale of decent usage" by being inconsistent, the suggestion to use them, merely because they maximize the "measure of rational belief when we are reasoning from the sample to the population," does not seem convincing.

However, are there cases where the maximum likelihood estimates are either inconsistent or inefficient? Yes, there are. The conditions where the maximum likelihood estimates are both consistent and efficient are stated in the papers by Hotelling, Doob and Wald quoted above. If these conditions are not satisfied, then, (i) the maximum likelihood estimates need not be consistent and, (ii) even if they are consistent, they need not be efficient. The following two examples demonstrating these possibilities are taken from the joint publication of Dr. E. L. Scott and myself.[11]

(i) Consider an increasing sequence of s series of measurements $x_{ij} (i = 1, 2, \cdots, s; j = 1, 2, \cdots, n_i)$. Assume that all the measurements are mutually independent and follow a normal law with the same variance σ^2. However, the quantity ξ_i measured in the ith series of measurements is different from the quantity ξ_j measured in the jth series. This is exactly the case where a fixed set of instruments is routinely used to measure different objects, perhaps characteristics of different stars. The joint probability density of all the observations is given by the formula

$$p_E \equiv \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\sum_i \sum_j (x_{ij}-\xi_i)^2/2\sigma^2}, \qquad N = \sum_{i=1}^{s} n_i.$$

In these circumstances it is frequently important to estimate σ, the standard error of measurements appropriate to the instruments used.

In Fisher's terminology, the likelihood function of a set of parameters means simply the probability density function (or a multiple of it) in which the particular values of the observable random variables are fixed and the parameters play the role of arguments. Thus, in the particular case considered, the likelihood function of the $s + 1$ parameters involved, namely, $\xi_1, \xi_2, \cdots, \xi_s$ and σ, is, say,

$$L = \text{const.} \times \sigma^{-N} e^{-\sum_i \sum_j (x_{ij}-\xi_i)^2/2\sigma^2}.$$

Given any system of observed values x_{ij} (for $i = 1, 2, \cdots, s; j = 1, 2, \cdots, n_i$) of the random variables X_{ij}, the maximum likelihood estimates of the

[11] J. Neyman and Elizabeth L. Scott: "Consistent estimates based on partially consistent observations." *Econometrica,* Vol. 16 (1948), pp. 1–32.

parameters are those values $\hat{\xi}_i$, $\hat{\sigma}$, for which L is a maximum. You will easily verify that the maximum likelihood estimates are

$$\hat{\xi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = x_{i\cdot} \qquad \text{for } i = 1, 2, \cdots, s$$

and

$$\hat{\sigma} = \left\{\frac{1}{N}\sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_{ij} - x_{i\cdot})^2\right\}^{\frac{1}{2}}.$$

We shall be particularly interested in the simplest case where $n_i = 2$ for $i = 1, 2, \cdots, s$. Then the square of the maximum likelihood estimate $\hat{\sigma}^2$ appears as a simple arithmetic mean,

$$\hat{\sigma}^2 = \frac{1}{s}\sum_{i=1}^{s}\left(\frac{1}{2}S_i^2\right)$$

of quantities

$$\tfrac{1}{2}S_i^2 = \tfrac{1}{2}[(x_{i1} - x_{i\cdot})^2 + (x_{i2} - x_{i\cdot})^2].$$

The expectation of this quantity is

$$E(\tfrac{1}{2}S_i^2) = \tfrac{1}{2}\sigma^2$$

and the variance, say,

$$V_{\frac{1}{2}S_i^2} = \frac{\sigma^4}{2}.$$

It follows that the variance of $\hat{\sigma}^2$ is

$$V_{\hat{\sigma}^2} = \frac{\sigma^4}{2s}$$

and tends to zero as s is increased. Thus, as s is indefinitely increased, $\hat{\sigma}^2$ tends in probability to its expectation $\sigma^2/2$ and, consequently, $\hat{\sigma}$ tends in probability not to σ but to the quantity $\sigma/\sqrt{2}$. It follows that, in this particular case, the maximum likelihood estimate of σ is inconsistent.

It may be said that the situation is trivial and that the bias in the estimate can be easily corrected by multiplying the estimate by $\sqrt{2}$. This is undoubtedly true but it is beside the point. It will be observed that the product $\hat{\sigma}\sqrt{2}$ is not the maximum likelihood estimate of σ and that the bias in $\hat{\sigma}$ does not tend to zero as the number s is increased. This is just the circumstance which the example is meant to illustrate.

(ii) In order to illustrate the possibility of the maximum likelihood estimate being consistent without being efficient, we shall use an example, similar to the above, where an increasing sequence of s series of measurements of the same quantity are made but where the error variance may vary from one series to the next. This is, for example, the case where ξ stands for the velocity

of light measured by s different observers, each using different equipment. As previously, we shall assume that all the measurements x_{ij}, for $i = 1, 2, \cdots, s$; $j = 1, 2, \cdots, n_i$, are independent and normally distributed about ξ so that their joint probability density function is

$$p \equiv \prod_{i=1}^{s} \left(\frac{1}{\sigma_i\sqrt{2\pi}}\right)^{n_i} e^{-\sum_{j=1}^{n_i} (x_{ij}-\xi)^2/2\sigma_i^2}.$$

As you will have no difficulty in verifying, the maximum likelihood estimate of ξ is the root of the equation, say

$$F_s(\hat{\xi}) \equiv \sum_{i=1}^{s} \frac{n_i(x_i. - \hat{\xi})}{S_i^2 + (x_i. - \hat{\xi})^2} = 0, \tag{18}$$

where S_i^2 has the usual meaning,

$$n_i S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - x_i.)^2.$$

The equation determining the maximum likelihood estimate $\hat{\xi}$ is complicated but can be solved numerically.

In addition to equation (18), the paper just quoted studies a more general equation, say,

$$\Phi_s(\tilde{\xi}) \equiv \sum_{i=1}^{s} \frac{w_i(x_i. - \tilde{\xi})}{S_i^2 + (x_i. - \tilde{\xi})^2} = 0, \tag{19}$$

obtained from (18) by substituting an arbitrary weight w_i for n_i. It is shown that, under mild restrictions regarding σ_i and w_i, the solution $\tilde{\xi}$ of (19) is a consistent estimate of ξ and that its variance is, say,

$$V_s(w) = \frac{1}{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2}} + \frac{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2}\left(\frac{w_i}{n_i - 2} - U\right)^2}{\left(\sum_{i=1}^{s} \frac{w_i}{\sigma_i^2}\right)^2},$$

where U stands for the weighted mean,

$$U = \frac{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2}\frac{w_i}{n_i - 2}}{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2}}.$$

It follows that the system of weights w_i which minimize the variance of the estimate $\tilde{\xi}$ are those for which

$$\frac{w_i}{n_i - 2} - U = 0, \qquad \text{for } i = 1, 2, \cdots, s,$$

or

$$w_i = (n_i - 2) \times \text{const.}$$

With these weights, equation (19) takes the form,

$$\sum_{i=1}^{s} \frac{(n_i - 2)(x_{i\cdot} - \tilde{\xi})}{S_i^2 + (x_{i\cdot} - \tilde{\xi})^2} = 0,$$

with the corresponding asymptotic variance of the solution equal to, say

$$V_{\text{opt.}} = \frac{1}{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2}}.$$

On the other hand, the asymptotic variance of the maximum likelihood solution is

$$V_{\hat{\xi}} = V_{\text{opt.}} + \frac{\sum_{i=1}^{s} \frac{n_i - 2}{\sigma_i^2} \left(\frac{n_i}{n_i - 2} - U \right)^2}{\left(\sum_{i=1}^{s} \frac{n_i}{\sigma_i^2} \right)^2}$$

and is, generally, greater than V_{opt}. The two variances V_{opt} and $V_{\hat{\xi}}$ coincide only if, for all $i = 1, 2, \cdots, s$, the number n_i of measurements forming the ith series is the same, say n. Then $U = n/(n - 2)$ and $V = V_{\text{opt}}$. Furthermore, as s is indefinitely increased, the quotient $V_{\text{opt}}/V_{\hat{\xi}}$ need not tend to unity so that the asymptotic efficiency of $\hat{\xi}$ is less than that of $\tilde{\xi}$. To see this you may wish to study more closely the simple particular case where $\sigma_1 = \sigma_2 = \cdots = \sigma_s = \sigma$ (though when estimating ξ we are not aware of this fact), and where $n_{2i-1} = n'$ and $n_{2i} = n''$ for all $i = 1, 2, \cdots, s$. It is also convenient to assume that s is an even number, say $s = 2m$. You will see that, in this particular case the quotient $V_{\text{opt}}/V_{\hat{\xi}}$ has a value independent of m and less than unity. Thus, it is shown that, even if the maximum likelihood estimate is consistent, it need not be efficient and that, on occasion, consistent and asymptotically normal estimates are easily constructed with variances smaller than that of the maximum likelihood estimate. In the next conference I shall attempt to show that smallness of the variance combined with consistency and asymptotic normality of an estimate means a substantial advantage in terms of consequences of the systematic use of a given estimation procedure. For me personally, this constitutes a decisive argument against the principle of maximum likelihood treated as a principle in the strict sense of the word. The following quotation from Fisher seems to suggest that this would also be his opinion.

In the present paper I have been particularly concerned to show that all the properties of mathematical likelihood, which make it valuable, can be demonstrated independently of any postulated value. From this it seems to me to follow that the concept of likelihood could be eliminated completely from discussions of estimation, and these discussions be adequately, though perhaps more cumbrously, carried out in other terms. (R. A. Fisher: "The logic of inductive inference." *Jr. Roy. Stat. Soc.,* Vol. 98 (1935), p. 81.)

However, on the next page of the same paper we read:

The fact that likelihood has been an aid to thought in such progress as has so far been made in the subject will suggest the advisability of using it for what it is worth, even though, ultimately, we may find ourselves able to do better. That there are logical situations in which the uncertainty of our inferences is expressible in terms of likelihood, but not in terms of probability, is one solid step gained, even though more comprehensive notions may later be developed. (R. A. Fisher: "The logic of inductive inference." *Jr. Roy. Stat. Soc.,* Vol. 98 (1935), p. 82.)

This passage requires comments from two different points of view. First, I wish to point out that Fisher's presumption that the mathematical concept of probability is inadequate when we are faced with the problem of estimation is based solely on his (quite correct) realization that, when *a priori* probabilities are not available (which he presumed to be always the case and which I agree is almost always the case), then the formula of Bayes is not applicable. Now, the general inadequacy of a concept is something which requires proof and the fact that one particular use of a concept is inapplicable does not, by itself, prove that no other uses of the same concept are possible with which to create an adequate basis for the theory of estimation. In fact, as you will see in the next conference, a theory of statistical estimation was developed entirely within the classical theory of probability, a theory which uses no other concepts and is applicable without any reference to probabilities *a priori.* For some questions which will be discussed later, it is relevant that Fisher, when writing his paper of 1935, was still of the opinion that the theory of probability by itself is not adequate for treating problems of estimation.

My other comment concerns the utility of the concept of likelihood as a measure of our confidence in the particular values of unknown parameters. If one attempts to answer the question "Why does Fisher think that in certain logical situations the likelihood is adequate to express our uncertainty?" one is forced to refer to passages like the last but one quoted. Here one finds contentions to the effect that the maximum likelihood estimates have properties which make the likelihood valuable and "which can be demonstrated independently of any postulated value." This demonstration appears to be on the ground of probability theory. Thus, the general argument is that the likelihood is an adequate measure of our confidence *because* the estimates obtained on this ground (or so Fisher thought) possess

certain desirable probabilistic properties. In these circumstances, the contention that the likelihood is adequate in cases where the concept of probability is not appears baseless and the references to *likelihood as a measure of confidence* contribute nothing but a certain amount of confusion. If Fisher's presumption that the desirable probabilistic properties (consistency and efficiency) are universally possessed by the maximum likelihood estimates were correct, then, except for this confusion of thought, the notion of the likelihood as a measure of confidence would not be harmful. However, as things stand, the notion of the new measure of confidence is regrettable because it may mislead the credulous part of the consumers of statistical theory.

All this applies to the notion of *likelihood as a measure of confidence.* On the other hand, there is no reason to object to the use of the label "likelihood function" applied to the probability density of the observable random variables with fixed particular values of these variables, considered as a function of the parameters.

To sum up: whenever the conditions of a particular problem imply that an unknown parameter is a random variable with a specified distribution *a priori,* then the formula of Bayes provides a clear cut solution of the problem of estimation; this solution, either in the form of a single estimate or in the form of an estimating interval, classical or modernized, has a simple interpretation in terms of frequencies of successes in estimating the unknown parameter; when the conditions of the practical problem considered do not imply the *a priori* distribution of the estimated parameter, then it is still "legitimate" to use the formula of Bayes; however, notwithstanding the theorem of Bernstein-v. Mises and the attempts to estimate the *a priori* distribution from past experience, such applications of Bayes' formula have a doubtful frequency interpretation; finally, it appears unprofitable to adopt the principle of maximum likelihood (and this also applies to the principle of insufficient reason and to its more recent modifications) because cases exist in which a strict adoption of this principle would lead to excessively frequent large errors in estimation that are perfectly avoidable.

Part 2. Outline of the Theory of Confidence Intervals

(Based on a conference held in the auditorium of the United States Department of Agriculture, April 9, 1937, 10 A.M., Dr. Frederick V. Waugh presiding.)

This morning I shall resume the outline of the problem of statistical estimation at the point where I stopped yesterday. You will remember that our discussion ended with the general conclusion that the classical approach by means of the theorem of Bayes provides a satisfactory solution only in the exceptional cases where the *a priori* distribution of the estimated param-

eter is known. In all other cases we have at best approximations of unknown precision, and at worst gross misconceptions dressed in impressive phraseology.

My purpose this morning is to explain a new method of approach to the problem of estimation especially designed for all cases in which the *a priori* distribution of the estimated parameter is not known and where, therefore, the estimated parameter may be treated as an unknown constant, not as a random variable. The theory I am going to present is known as the theory of confidence intervals. The first outline of this theory appeared in my paper [1] of 1934. A more thorough treatment is found in two subsequent memoirs, one in English [2] and the other in French.[3] However, the first reference to confidence intervals appeared in 1932 in a monograph [4] of Wacław Pytkowski, then a student of mine, who applied the new theory to the problems of estimation of various characteristics of small farms in Poland.

When approaching the practical problem of estimation in cases where no information about the *a priori* distribution is available, it is important to realize that the corresponding mathematical problem should be stated in a form which is essentially different from the form leading to Bayes' solution. The Bayes' solution answers the following question: (a) *given that the observable random variables* $X_1, X_2, \cdots, X_n$ *have assumed the specified values* $x_1, x_2, \cdots, x_n$, *what is the probability,*

$$P\{a < \theta_1 \leqq b \mid (X_1 = x_1)(X_2 = x_2) \cdots (X_n = x_n)\},$$

that the estimated parameter θ_1 *will have a value contained between the specified limits* $a < b$? As we have seen, the answer to this question depends upon the *a priori* distribution of θ_1 and, if this distribution is not known, question (a) cannot be answered. Thus, if a solution of the practical problem of estimation is to be based on the theory of probability, it will be necessary to formulate a new problem, say (b), different from (a). Problem (b) must be such that its solution will not depend upon the *a priori* distribution of the estimated parameter and, at the same time, will give an

[1] J. Neyman: "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." *Jr. Roy. Stat. Soc.,* Vol. 97 (1934), pp. 558–625.

[2] J. Neyman: "Outline of a theory of statistical estimation based on the classical theory of probability." *Phil. Trans. Roy. Soc., London,* Ser. A, Vol. 236 (1937), pp. 333–380.

[3] J. Neyman: "L'estimation statistique traitée comme un problème classique de probabilité." *Actualités Scientifiques et Industrielles,* No. 739 (1938), pp. 25–57.

[4] Wacław Pytkowski: "The dependence of the income in small farms upon their area, the outlay and the capital invested in cows." *Bibljoteka Puławska,* No. 34 (1932), Warszawa, 59 pp. + 4 tables.

intelligible answer to the difficulty facing the practical statistician. It appears that both the formulation of the new probabilistic problem (b) and its solution are very simple and the fact that they were not found for a long time must be ascribed to what Karl Pearson called "routine of thought" and to attachment to the formula of Bayes. The scholars must have been so impressed by Bayes' formula that they just did not think of thinking about the problem in a different manner. However, as we shall see later, the elements of the new idea can be discovered in the writings of many earlier authors, beginning with Gauss. Unfortunately, these elements of thought, for some reason, took quite a long time to grow and to crystallize.

We shall begin by recalling what exactly the practical statistician does when he is faced with a problem of estimation and what exactly he needs from the theory. In this, our attention will be primarily directed towards the problem of estimation by interval and we shall have to return to the ideas described in the early part of yesterday's conference.

We contemplate a situation in which the practical statistician is interested in the value of the parameter θ_1 that appears in the probability density function $p_E(x_1, x_2, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_s)$ of n observable random variables $X_1, X_2, \cdots, X_n$. The analytical form of this probability density function is known to the statistician, but the values of the parameters $\theta_1, \theta_2, \cdots, \theta_s$ are unknown, except that they are contained in some specified intervals, say $A_i < \theta_i < B_i$ $(i = 1, 2, \cdots, s)$, finite or infinite. The practical statistician is faced with the necessity of taking an action which should be adjusted to the value of the parameter θ_1. If he is the M.D. of the second example in yesterday's conference, the action contemplated consists in administering to the patient a dose of drug B, a dose which should be appropriately adjusted to the content η of substance A in the patient's blood. If the practical statistician is concerned with the policy of the Department of Agriculture, his contemplated action may consist in suggesting a provision in a forthcoming bill, a provision which should be adjusted to the characteristic ξ of the totality of farms in the United States as mentioned in the first example of the last conference. Unfortunately, neither ξ nor η can be evaluated exactly and the best the two practical statisticians can do is to observe particular values of the random variables $X_1, X_2, \cdots, X_n$ and base their actions on these observations. In each case the assertion about the true value of the unknown parameter θ_1 (ξ in one case and η in the other) will be made in the same form,

$$\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq \theta_1 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n), \tag{1}$$

where $\underline{\theta}$ and $\bar{\theta}$ are functions of the observable random variables. Then the practical statistician will adjust his actions as if it were known for certain that the true value of θ_1 is contained between the limits indicated.

All human actions are subject to error and the actions of the practical statistician cannot be an exception to the general rule. Thus the practical statistician must be aware that, whatever function $\underline{\theta}$ and $\bar{\theta}$ he selects, his assertions about the value of θ_1 will be erroneous from time to time. The best he can hope to arrange is that the errors of estimation do not occur too frequently. Also, he may have in mind a scale of importance of different errors. For example, in a particular case an overestimate of the parameter may be more important to avoid than an underestimate. Finally, the practical statistician is likely to desire that the difference $\bar{\theta} - \underline{\theta}$ be, in general, as small as possible. However, the most pressing need which the practical statistician is likely to feel is that he be given the opportunity to select a number α, $0 < \alpha < 1$, just as close to unity as he desires, and to determine a pair of functions $\underline{\theta}(X_1, X_2, \cdots, X_n)$ and $\bar{\theta}(X_1, X_2, \cdots, X_n)$ such that their use to estimate θ_1 in the manner described will yield correct results with the long-run relative frequency equal to α or, if this is impossible, at least equal to α.

As a general result of this discussion, we can now formulate the mathematical problem (b) referring to the problem of estimating a parameter θ_1 which in the modern form of theory of estimation takes the place of problem (a) discussed above.

Problem (b). *Given that the observable random variables* $E \equiv (X_1, X_2, \cdots, X_n)$ *follow a distribution with the probability density function* $p_E(x_1, x_2, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_s)$ *depending on* s *parameters* $\theta_1, \theta_2, \cdots, \theta_s$, *the values of which are unknown; given also that the parameter* θ_i *may have any value between the specified limits* $A_i < \theta_i < B_i$ *for* $i = 1, 2, \cdots, s$; *finally, given a number* α *between the limits* $0 < \alpha < 1$, *to determine two functions* $\underline{\theta}(x_1, x_2, \cdots, x_n)$ *and* $\bar{\theta}(x_1, x_2, \cdots, x_n)$ *defined over all possible systems of values* $x_1, x_2, \cdots, x_n$ *of the observable random variables, such that for all possible systems of values of the parameters* $\theta_1, \theta_2, \cdots, \theta_s$

$$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq \theta_1 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n) \mid \theta_1, \theta_2, \cdots, \theta_s\} \equiv \alpha. \quad (2)$$

It is essential to be entirely clear about the implications of the requirements imposed on the two functions $\underline{\theta}$ and $\bar{\theta}$. You will notice the sign of identity $\equiv$ appearing in formula (2). This sign emphasizes the requirement that the probability on the left be equal to α irrespective of what value θ_1 takes between $A_1 < \theta_1 < B_1$, and irrespective of the values of the other parameters $\theta_2, \theta_3, \cdots, \theta_s$. Thus, in particular, if unity, two and three are the possible values of θ_1, it is required from the functions $\underline{\theta}$ and $\bar{\theta}$ that

$$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq 1 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n) \mid (\theta_1 = 1), \theta_2, \cdots, \theta_s\} \equiv \alpha,$$

$$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq 2 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n) \mid (\theta_1 = 2), \theta_2, \cdots, \theta_s\} \equiv \alpha,$$

$$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq 3 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n) \mid (\theta_1 = 3), \theta_2, \cdots, \theta_s\} \equiv \alpha,$$

etc., where the identity signs refer to the possibility of variation in the values of $\theta_2, \theta_3, \cdots, \theta_s$ and require that the probability on the left hand side keeps the same value α, irrespective of changes in the values of $\theta_2, \theta_3, \cdots, \theta_s$. In other words, it is required of the functions $\underline{\theta}$ and $\bar{\theta}$ that, if $\theta_1 = 1$, they bracket unity with the prescribed frequency α. If $\theta_1 = 2$, they are required to bracket 2 with the same frequency α, etc.

It is seen that our requirements regarding the functions $\underline{\theta}$ and $\bar{\theta}$ are tricky and, at least at first sight, one does not know quite where to begin to satisfy them. However, it was possible to prove that the problem of determining $\underline{\theta}$ and $\bar{\theta}$ reduces to another problem of a more familiar nature. Now let us adopt the following definition. We denote by α a fixed number between zero and unity.

If two functions $\underline{\theta}(X_1, X_2, \cdots, X_n)$ and $\bar{\theta}(X_1, X_2, \cdots, X_n)$ of the observable random variables $X_1, X_2, \cdots, X_n$, defined over all possible systems of values of these variables, possess the property that, whatever the possible value of the parameter θ_1, the probability (2) *of $\underline{\theta}(X_1, X_2, \cdots, X_n)$ falling short of θ_1 at the same time that $\bar{\theta}(X_1, X_2, \cdots, X_n)$ is at least equal to θ_1 equals α identically in $\theta_2, \theta_3, \cdots, \theta_s$, then we shall say that $\underline{\theta}$ and $\bar{\theta}$ are the lower and the upper confidence limits for θ_1 corresponding to the confidence coefficient α.*

Furthermore, *the interval $[\underline{\theta}(X_1, X_2, \cdots, X_n), \bar{\theta}(X_1, X_2, \cdots, X_n)]$ will be called the confidence interval for θ_1 corresponding to the confidence coefficient α.*

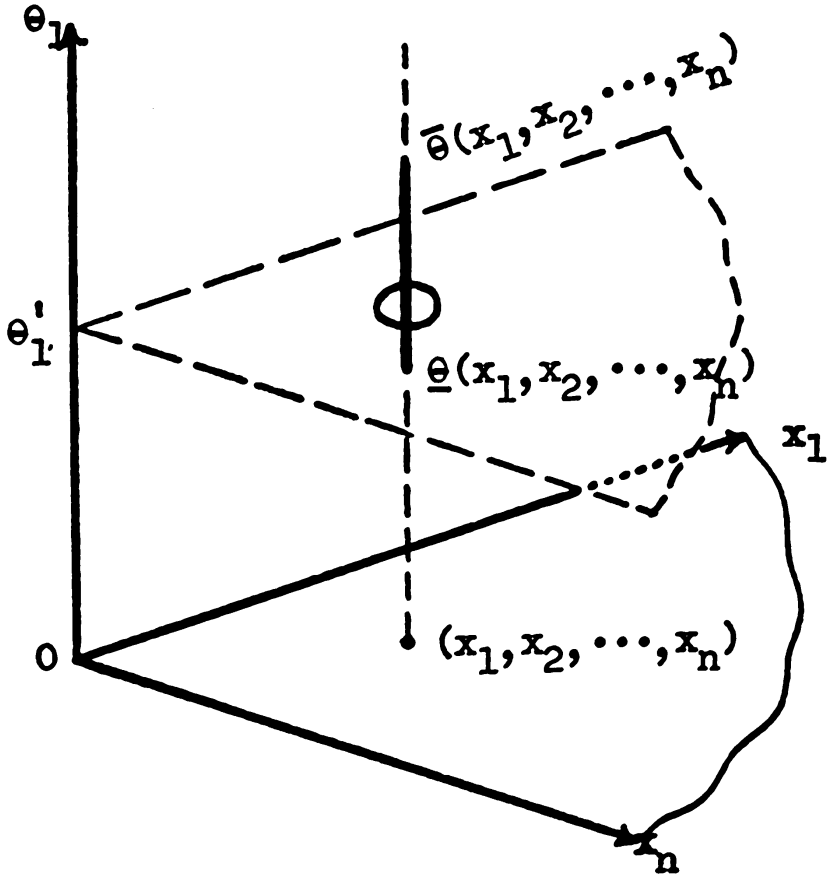
In an earlier part of this book, we have used the terms *sample point* and *sample space*. If $x_1, x_2, \cdots, x_n$ are possible values of the observable random variables $X_1, X_2, \cdots, X_n$, then we say that the system of n numbers $(x_1, x_2, \cdots, x_n)$ determines (or represents) a *possible sample point*. The set of all possible sample points is called the *sample space* and is denoted by W. If n does not exceed 3, then the sample points and the sample space are easily interpreted in the space of the appropriate number of dimensions and are easy to visualize. If n is greater than 3, diagrammatic presentation is impossible but it is still convenient to speak in terms of points and spaces.

We shall now indicate how the search for confidence limits is reduced to the search for certain regions in the sample space, called regions of acceptance. For this purpose assume for a moment that the confidence limits $\underline{\theta}$ and $\bar{\theta}$ have already been found and correspond to a confidence coefficient α, previously selected, $0 < \alpha < 1$. Consider a space G (general space) of $n + 1$ dimensions. Of the $n + 1$ axes of coordinates in this space, the first n will correspond to the n observable random variables. In other words, the possible values of X_1 will be measured on axis Ox_1, the possible values of X_2 will be measured on axis Ox_2, etc. On the last (the $n + 1$-st) axis of coordinates in G, we shall measure the possible values of the estimated parameter θ_1. We shall imagine that this last axis $O\theta_1$ is vertical. Now, select any possible sample point $(x_1, x_2, \cdots, x_n)$. To this point, there will correspond a

value $\underline{\theta}(x_1, x_2, \cdots, x_n)$ of the lower confidence limit and a value $\bar{\theta}(x_1, x_2, \cdots, x_n)$ of the upper confidence limit. Imagine that we plot the two points, $[x_1, x_2, \cdots, x_n, \underline{\theta}(x_1, x_2, \cdots, x_n)]$ and $[x_1, x_2, \cdots, x_n, \bar{\theta}(x_1, x_2, \cdots, x_n)]$, and connect them by a line. This line or, rather, this interval of line, call it $\delta(x_1, x_2, \cdots, x_n)$, will represent the confidence interval corresponding to the

FIGURE 1

General space and confidence intervals



selected possible sample point. The situation is illustrated in Figure 1. Imagine that this procedure is repeated for each and every possible sample point. Now, take a possible value of θ_1, say θ_1', and, in the general space G, consider a horizontal plane $\theta_1 = \theta_1'$. Generally, this plane will cut some of the confidence intervals and will miss others.

Denote by $A(\theta_1')$ the set of all possible sample points such that the corresponding confidence intervals are cut by the plane $\theta_1 = \theta_1'$. In other words, the set $A(\theta_1')$ is the set of all possible sample points that satisfy the double condition,

$$\underline{\theta}(x_1, x_2, \cdots, x_n) \leqq \theta_1' \leqq \bar{\theta}(x_1, x_2, \cdots, x_n). \tag{3}$$

The set $A(\theta_1')$ so defined is called the *region of acceptance* corresponding to θ_1'. Thus, if $\underline{\theta}$ and $\bar{\theta}$ are confidence limits for θ_1 corresponding to the confidence coefficient α, then they determine a set, say A, of regions of acceptance. Obviously, to each possible value of θ_1 there corresponds a region of acceptance.

The regions of acceptance $A(\theta_1)$ and their set A possess the following properties. The important property of every particular region of acceptance, say $A(\theta_1')$, is that if a possible sample point $(x_1, x_2, \cdots, x_n)$ falls within $A(\theta_1')$, then the corresponding confidence interval $\delta(x_1, x_2, \cdots, x_n)$ covers the value θ_1' of θ_1 and vice versa. This is an immediate consequence of the definition of $A(\theta_1')$ by means of the double relation (3). In fact, in order to verify whether or not a sample point $(x_1', x_2', \cdots, x_n')$ falls within $A(\theta_1')$, it is sufficient to compute the values of $\underline{\theta}$ and $\bar{\theta}$ corresponding to this point and see whether or not they bracket θ_1'. But this is exactly what we would do in order to verify whether or not $\delta(x_1', x_2', \cdots, x_n')$ covers θ_1'. In order to express this by a formula, we shall agree to use the letter C to denote the word "covers" and the letter ϵ to denote the phrase "is an element of" or "belongs to." With this notation, we may write the identity of the two events,

$$[E \epsilon A(\theta_1')] \equiv [\delta(E) \text{ C } \theta_1'], \tag{4}$$

where, as formerly, the letter E stands for the set of observable random variables $X_1, X_2, \cdots, X_n$.

It follows from (4) that, whatever be the assumptions on which the probabilities are computed,

$$P\{E \epsilon A(\theta_1')\} \equiv P\{\delta(E) \text{ C } \theta_1'\}. \tag{5}$$

In particular, if we compute the probabilities on the assumption that $\theta_1 = \theta_1'$ while the other parameters $\theta_2, \theta_3, \cdots, \theta_s$ have arbitrary values, we shall find

$$P\{E \epsilon A(\theta_1') \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv P\{\delta(E) \text{ C } \theta_1' \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv \alpha,$$

because of the definition of confidence intervals. Thus, if θ_1' is a possible value of θ_1 and $A(\theta_1')$ is the corresponding region of acceptance, then, whatever be the possible values of $\theta_2, \theta_3, \cdots, \theta_s$,

$$P\{E \epsilon A(\theta_1') \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv \alpha. \tag{6}$$

Identity (6) represents the necessary condition which a region $A(\theta_1')$ must satisfy in order to qualify as a region of acceptance corresponding to the value θ_1' of the parameter θ_1. In addition to this condition which applies to each and every region of acceptance $A(\theta_1')$ taken separately, there are important conditions which apply to the whole set A of regions of acceptance. These conditions are intuitive and, therefore, I am going to enumerate them without giving proofs. The proofs are given in detail in my paper of 1937, already quoted.

(I) The first condition that the set A of regions of acceptance must satisfy is that the union of all regions of acceptance, corresponding to all possible values of the parameter θ_1, must coincide with the sample space W. In other words, whatever possible sample point we take, say $(x_1, x_2, \cdots, x_n)$, there must exist at least one possible value of θ_1 such that its region of acceptance includes this particular sample point.

(II) The second necessary condition which the set A must satisfy is as follows. Let $(x_1', x_2', \cdots, x_n')$ be an arbitrary possible sample point. According to the above condition (I), there exists at least one possible value θ_1' of θ_1 such that $A(\theta_1')$ contains $(x_1', x_2', \cdots, x_n')$. Consider the set $S(x_1', x_2', \cdots, x_n')$ of all possible values of θ_1 such that the point $(x_1', x_2', \cdots, x_n')$ is contained within their regions of acceptance. Then condition (II) states that the set $S(x_1', x_2', \cdots, x_n')$ fills *a closed interval.* This closed interval extends from $\underline{\theta}(x_1', x_2', \cdots, x_n')$ to $\bar{\theta}(x_1', x_2', \cdots, x_n')$.

It is easy to see that condition (6), applying to each region $A(\theta_1')$ separately, and conditions (I) and (II), applying to the set A of all regions $A(\theta_1')$, are necessary and sufficient for the set A to be the set of regions of acceptance. In other words, if we start with defining for each possible value θ_1' of θ_1 a region $A(\theta_1')$ satisfying (6) and if we manage to adjust these regions so that their set A satisfies conditions (I) and (II), then this set A determines confidence intervals for θ_1 corresponding to the confidence coefficient α. In fact, suppose that (6) and (I) and (II) are satisfied by some regions, say $B(\theta_1')$. Let $(x_1, x_2, \cdots, x_n)$ be a possible sample point. According to (I) and (II), there exists a closed interval of possible values of θ_1, extending from some value $f_1(x_1, x_2, \cdots, x_n)$ to some other value $f_2(x_1, x_2, \cdots, x_n)$ such that, whatever θ_1'' between the limits, $f_1 \leqq \theta_1'' \leqq f_2$, the point $(x_1, x_2, \cdots, x_n)$ belongs to the region $B(\theta_1'')$. You will have no difficulty in verifying that the two functions f_1 and f_2 satisfy the definition of confidence limits for θ_1. In fact, the interval between them, say $\Delta(x_1, x_2, \cdots, x_n)$, covers any given value θ_1' of θ_1 whenever $(x_1, x_2, \cdots, x_n)$ belongs to $B(\theta_1')$ and in no other case. On the other hand, since $B(\theta_1')$ is supposed to satisfy condition (6), we have

$$P\{E \epsilon B(\theta_1') \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv \alpha.$$

Thus,

$$P\{\Delta(E) \subset \theta_1' \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv P\{E \epsilon B(\theta_1') \mid \theta_1', \theta_2, \cdots, \theta_s\} \equiv \alpha.$$

In this way we come to the conclusion that, in order to determine a pair of confidence limits, it is both necessary and sufficient to determine a family A of regions of acceptance satisfying the above three conditions.

If the number s of unknown parameters involved in the probability density function of the observable random variables exceeds unity, there are substantial difficulties in determining regions satisfying condition (6). Regions satisfying this condition are called *similar to the sample space* and there is

a substantial literature concerning them.[5] On the other hand, if $s = 1$, the problem of satisfying condition (6) is trivial. Conditions (I) and (II) are also easy to satisfy and, generally, we have at our disposal a great variety of confidence intervals corresponding to the same confidence coefficient. In the simplest case, $s = 1$, the theory of confidence intervals is concerned mainly with the problem of an appropriate choice among the many confidence intervals. Naturally, in the case $s > 1$, the problem of choice also exists and presents more difficulties. Limitations of time and space make it impossible to discuss all of these matters in detail and I must refer you to the literature already quoted. The best that I can do here is to work out an example in order to illustrate the procedure of determining confidence intervals. In so doing, we shall have occasion to discuss the interpretation of confidence intervals and also to touch upon the problem of optimum.

The example I shall use is the one discussed yesterday. This is the example of n observable random variables $X_1, X_2, \cdots, X_n$, all independent and having the same distribution with the probability density function equal to $1/\theta$ for $0 < x \leqq \theta$ and zero elsewhere, $\theta > 0$ being the parameter to be estimated. Yesterday I considered the particular case in which some definite information regarding θ was available. Namely, I assumed as known for certain that (a) θ cannot exceed the limits $0 < \theta \leqq 1$ and (b) the frequency of cases where θ falls within any interval (a, b) partial to $(0, 1)$ is represented by the integral,

$$\int_a^b m\theta^{m-1}\, d\theta = b^m - a^m,$$

with a known value of m. In other words, I assumed in discussing this example that the *a priori* distribution of θ was known exactly.

Today, contrary to this, I shall study the problem of estimating θ in conditions where nothing whatever is known about its value except that it

[5] See, for example, the following papers:

(a) J. Neyman and E. S. Pearson: "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc., London,* Ser. A, Vol. 231 (1933), pp. 289–337.

(b) W. Feller: "Note on regions similar to the sample space." *Stat. Research Memoirs,* Vol. 2 (1938), pp. 117–125.

(c) J. Neyman: "On a statistical problem arising in routine analysis and in sampling inspection of mass production." *Annals of Math. Stat.,* Vol. 12 (1941), pp. 46–76.

(d) H. Scheffé: "On the theory of testing composite hypotheses with one constraint." *Annals of Math. Stat.,* Vol. 13 (1942), pp. 280–293.

(e) E. L. Lehmann: "On optimum tests of composite hypotheses with one constraint." *Annals of Math. Stat.,* Vol. 18 (1947), pp. 473–493.

(f) P. G. Hoel: "On the uniqueness of similar regions." *Annals of Math. Stat.,* Vol. 19 (1948), pp. 66–71.

is a positive number. Of course, this unique assumption cannot be considered as any sort of limitation to the generality of the problem, since the assumption $\theta > 0$ is implied by the nature of the assumed distribution of the observable random variables, with their probability density function positive within the interval $(0, \theta)$ and zero elsewhere.

Thus, the difference in the conditions of the problem of estimation considered yesterday and this morning is as follows: Yesterday I assumed definite information regarding the distribution of the observable random variables given the value of the parameter θ *and* definite information regarding the distribution of θ considered as a random variable; this morning I shall treat the problem of estimating θ when no assumptions are made regarding the *a priori* distribution of θ except the one implied by the information about the distribution of the observable random variables. Our problem will be to construct a system of confidence intervals for θ corresponding to a preassigned confidence coefficient α, say $\alpha = .90$, $\alpha = .95$, etc.

Turning to the general theory outlined above, we must be clear about our aims and about the steps we have to take to attain these aims. Our aim is to define over the whole sample space W of the observable random variables two functions $\underline{\theta}(X_1, X_2, \cdots, X_n)$ and $\bar{\theta}(X_1, X_2, \cdots, X_n)$ having the property that, whatever be the (necessarily positive) value θ_1 of the parameter θ, the probability that this value θ_1 will be bracketed by $\underline{\theta}$ and $\bar{\theta}$ is equal to α,

$$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) \leqq \theta_1 \leqq \bar{\theta}(X_1, X_2, \cdots, X_n) \mid \theta = \theta_1\} \equiv \alpha. \quad (7)$$

This, then, is our aim. The means to attain this aim, as indicated by the foregoing theory, is to take the following steps:

(i) To determine the sample space W;

(ii) For each possible value θ_1 of θ, i.e. for each positive number θ_1, to select within W a region of acceptance $A(\theta_1)$ satisfying condition (6) and such that the totality A of such regions satisfy conditions (I) and (II).

Then the boundaries of the set $S(x_1, x_2, \cdots, x_n)$ will represent the values of the functions $\underline{\theta}$ and $\bar{\theta}$ corresponding to the possible sample point $(x_1, x_2, \cdots, x_n)$.

I have already pointed out that in many cases not one but many different systems of regions of acceptance are available. Naturally, each system of regions of acceptance determines a separate system of confidence intervals corresponding to the same confidence coefficient α. In order to illustrate this point, we shall select two systems of regions of acceptance, A and B, and examine the corresponding confidence intervals.

According to the conditions of the problem, the probability density function of the n observable random variables, $X_1, X_2, \cdots, X_n$, is given by

$$p_E(x_1, x_2, \cdots, x_n \mid \theta) = \frac{1}{\theta^n} \qquad \text{for } 0 < x_1, x_2, \cdots, x_n \leqq \theta$$
$$= 0 \qquad \text{elsewhere,} \tag{8}$$

where θ is some unknown positive number. Thus, function (8) is positive within a hypercube of n dimensions, with each dimension extending from zero to θ. Since $\theta > 0$ may be any number, every point with positive coordinates $x_1, x_2, \cdots, x_n$ is a possible sample point. It follows that, in this case, the sample space W is the set of all points in n dimensions with their coordinates $x_i > 0$, for $i = 1, 2, \cdots, n$. This statement completes step (i).

Now, let us proceed to step (ii) and determine the system A of regions of acceptance. For this purpose, fix for a moment an arbitrary possible value θ_1 of θ, and denote by $W(\theta_1)$ the region partial to W determined by the inequalities,

$$0 < x_i \leqq \theta_1, \qquad \text{for } i = 1, 2, \cdots, n. \tag{9}$$

Should it happen that $\theta_1 > 0$ is the true value of θ, then within $W(\theta_1)$ the probability density function (8) of the observable random variables is positive and equal to $1/\theta_1^n$, while outside of $W(\theta_1)$ it is zero. With the notation adopted, the symbol $W[\theta_1(1 - \alpha)^{1/n}]$ denotes a hypercube partial to $W(\theta_1)$ with dimensions,

$$0 < x_i \leqq \theta_1(1 - \alpha)^{1/n}, \qquad \text{for } i = 1, 2, \cdots, n. \tag{10}$$

As the region of acceptance, $A(\theta_1)$, corresponding to the selected possible value θ_1 of θ, we shall select that part of the hypercube $W(\theta_1)$ which lies outside of $W[\theta_1(1 - \alpha)^{1/n}]$, with the inclusion of the outer boundary of the latter. In other words, $A(\theta_1)$ is defined to include every point $(x_1, x_2, \cdots, x_n)$ which satisfies condition (9) but fails to satisfy the condition

$$0 < x_i < \theta_1(1 - \alpha)^{1/n}, \qquad \text{for } i = 1, 2, \cdots, n. \tag{11}$$

You will notice that (11) differs from (10) by the lack of the equality sign of the right.

It is easy to see that region $A(\theta_1)$ satisfies condition (6). To see this, assume that θ_1 is the true value of θ and compute the probability $P\{E \in A(\theta_1) \mid \theta_1\}$. Owing to the fact that the distribution of E within $W(\theta_1)$ is uniform on the assumption just made, the probability, $P\{E \in A(\theta_1) \mid \theta_1\}$ is equal to the volume of $A(\theta_1)$ divided by the volume of $W(\theta_1)$. According to the definition of $A(\theta_1)$,

$$\begin{aligned} \text{Volume of } A(\theta_1) &= \text{Volume of } W(\theta_1) - \text{Volume of } W[\theta_1(1 - \alpha)^{1/n}] \\ &= \theta_1^n - [\theta_1(1 - \alpha)^{1/n}]^n \\ &= \alpha\theta_1^n \end{aligned}$$

and it follows that

$$P\{E \epsilon A(\theta_1) \mid \theta_1\} = \alpha$$

irrespective of the value θ_1 of θ.

Thus, for every possible value θ_1 of θ, we have defined a region $A(\theta_1)$ satisfying condition (6). In order to determine whether or not these regions can be regarded as regions of acceptance, we shall consider the set A of all these regions and verify whether or not it satisfies conditions (I) and (II). For this purpose it will be convenient to give the definition of $A(\theta)$ an analytical form. You will notice that, in order to determine whether a given point $(x_1, x_2, \cdots, x_n)$ with positive coordinates falls within $A(\theta)$ or not, it is not necessary to know the values of all of its coordinates. For this particular purpose, it is sufficient to know the value x of the greatest of the n coordinates $x_1, x_2, \cdots, x_n$. If

$$\theta(1 - \alpha)^{1/n} \leqq x \leqq \theta, \tag{12}$$

then the point $(x_1, x_2, \cdots, x_n)$ belongs to $A(\theta)$. Otherwise, it does not. Thus, the double formula (12) represents the complete definition of the region $A(\theta)$ corresponding to any specified $\theta > 0$.

In order to verify that the set A satisfies conditions (I) and (II), fix an arbitrary possible sample point, i.e., a point with arbitrary positive coordinates $x_1, x_2, \cdots, x_n$, and determine the set $S(x_1, x_2, \cdots, x_n)$ of values of θ for which this point $(x_1, x_2, \cdots, x_n) \epsilon A(\theta)$. As previously, let x denote the greatest of the coordinates of the selected possible sample point. In order that this point belong to $A(\theta)$, it is necessary and sufficient that θ satisfy the double condition (12). Solving this condition for θ, we obtain the double condition

$$x \leqq \theta \leqq \frac{x}{(1 - \alpha)^{1/n}} \tag{13}$$

which defines the set $S(x_1, x_2, \cdots, x_n)$. It is seen that the set $S(x_1, x_2, \cdots, x_n)$ extends over a closed interval beginning with x and ending with $x/(1 - \alpha)^{1/n}$.

It follows that the set A of regions $A(\theta)$ satisfies conditions (I) and (II) and that the corresponding confidence limits for θ are

$$\underline{\theta}(x_1, x_2, \cdots, x_n) = x,$$

and

$$\bar{\theta}(x_1, x_2, \cdots, x_n) = \frac{x}{(1 - \alpha)^{1/n}}.$$

The length of the confidence interval corresponding to the given point $(x_1, x_2, \cdots, x_n)$ is, say,

$$\delta(x_1, x_2, \cdots, x_n) = x \frac{1 - (1 - \alpha)^{1/n}}{(1 - \alpha)^{1/n}}.$$

Let X be defined as the random variable whose value coincides with that of the greatest of the random variables $X_1, X_2, \cdots, X_n$. Then the substitution in $\underline{\theta}$ and $\bar{\theta}$ of X_i instead of x_i (for $i = 1, 2, \cdots, n$) and of X instead of x, will yield two random variables,

$$\begin{aligned} \underline{\theta}(X_1, X_2, \cdots, X_n) &= X, \\ \bar{\theta}(X_1, X_2, \cdots, X_n) &= \frac{X}{(1 - \alpha)^{1/n}}, \end{aligned} \tag{14}$$

which have property (7) for all values of $\theta > 0$. In other words, whatever may be the true value of $\theta > 0$, the probability that the two functions (14) will bracket this value is exactly equal to α. Thus, the practical statistician who makes a rule of asserting that θ is a number between the particular values of $\underline{\theta}$ and $\bar{\theta}$ as determined by observation is in a position exactly comparable to that of a gambler betting on the outcome of a game of chance, the probability of which is equal to α which may be as close to unity as desired.

We shall use the symbol $\delta(E)$ to denote the confidence interval determined by the two functions (14) and built by using system A of regions of acceptance. Now, we shall proceed to define an alternative system B of regions of acceptance and the corresponding confidence interval, say $\Delta(E)$.

Fix a tentative positive value θ_1 of θ and define $B(\theta_1)$ to include all points of the sample space W such that the arithmetic mean $\bar{x}$ of their coordinates differs from $\theta_1/2$ by not more than a quantity $u(\theta_1)$. Thus, $B(\theta_1)$ is defined by the double relation,

$$\tfrac{1}{2}\theta_1 - u(\theta_1) \leqq \bar{x} \leqq \tfrac{1}{2}\theta_1 + u(\theta_1). \tag{15}$$

Let $\overline{X}$ represent the random variable defined as the arithmetic mean of $X_1, X_2, \cdots, X_n$. The motivation for the above choice of the region $B(\theta_1)$ is that $\theta_1/2$ represents the expected value of $\overline{X}$ and also its most probable value. The quantity $u(\theta_1)$ must be so determined as to satisfy condition (6). Since a given sample point does or does not belong to $B(\theta_1)$ according as $\overline{X}$ does or does not satisfy condition (15), it is obvious that

$$P\{E \in B(\theta_1) \mid \theta_1\} \equiv P\{\left| \overline{X} - \tfrac{1}{2}\theta_1 \right| \leqq u(\theta_1) \mid \theta_1\}.$$

It follows that the value of $u(\theta_1)$ can be found by using the probability density function of $\overline{X}$ computed on the assumption that θ_1 is the true value of θ. The exact form of this probability density function was found by Laplace. If $n = 2$, then it is very simple,

$$p_{\bar{X}}(\bar{x} \mid \theta_1) = 4\bar{x} \qquad \text{for } 0 \leqq \bar{x} \leqq \tfrac{1}{2}\theta_1,$$
$$= 4(\theta_1 - \bar{x}) \qquad \text{for } \tfrac{1}{2}\theta_1 \leqq \bar{x} \leqq \theta_1,$$
$$= 0 \qquad \text{elsewhere.}$$

On the other hand, as n is increased, the expression for $p_{\bar{X}}(\bar{x} \mid \theta_1)$ becomes more and more complicated and soon becomes unmanageable. It happens, however, that with very moderate values of n, say with $n \geqq 5$, the probabilities computed using the true probability density function are already difficult to distinguish from their normal approximations. Since our purpose here is to deal with the conceptual, rather than with the numerical side of the problem, we shall use the normal approximation of the probability density of $\overline{X}$. Using the fact that, for each observable random variable X_i,

$$E(X_i \mid \theta_1) = \tfrac{1}{2}\theta_1,$$
$$E(X_i^2 \mid \theta_1) = \tfrac{1}{3}\theta_1^2,$$

we find that

$$\sigma_{X_i}^2 = \tfrac{1}{12}\theta_1^2$$

and it follows that

$$E(\overline{X} \mid \theta_1) = \tfrac{1}{2}\theta_1,$$
$$\sigma_{\bar{X}}^2 = \frac{\theta_1}{\sqrt{12n}}.$$

The normal approximation to the probability density function $p_{\bar{X}}(\bar{x} \mid \theta_1)$ is, then, say

$$p_{\bar{X}}^*(\bar{x} \mid \theta_1) = \frac{\sqrt{12n}}{\theta_1\sqrt{2\pi}}\, e^{-12n(\bar{x}-\theta_1/2)^2/2\theta_1^2},$$

and the probability that $\overline{X}$ will differ from $\theta_1/2$ by not more than $u(\theta_1)$ is approximately equal to, say,

$$P^*\{\,|\,\overline{X} - \tfrac{1}{2}\theta_1\,|\, \leqq u(\theta_1)\} = 2\int_0^{u(\theta_1)} p_{\bar{X}}^*(\bar{x} \mid \theta_1)\, d\bar{x}. \tag{16}$$

After some easy algebra, this formula reduces to

$$P^*\{\,|\,\overline{X} - \tfrac{1}{2}\theta_1\,|\, \leqq u(\theta_1)\} = \sqrt{\frac{2}{\pi}}\int_0^{\lambda(\alpha)} e^{-t^2/2}\, dt,$$

where

$$\lambda(\alpha) = \frac{u(\theta_1)}{\theta_1}\sqrt{12n}. \tag{17}$$

The requirement that the probability (16) equal the preassigned value α, less than unity, determines uniquely the value of $\lambda(\alpha)$ which can be found in

any table of the normal integral. For example, if $\alpha = .95$, then $\lambda(\alpha) = 1.96$, etc. Now equation (17) determines $u(\theta_1)$, namely,

$$u(\theta_1) = \theta_1 \frac{\lambda(\alpha)}{\sqrt{12n}}$$

and it follows that, if we grant the normal approximation, region $B(\theta_1)$ is defined by the double relation,

$$\theta_1\left(\frac{1}{2} - \frac{\lambda(\alpha)}{\sqrt{12n}}\right) \leqq \bar{x} \leqq \theta_1\left(\frac{1}{2} + \frac{\lambda(\alpha)}{\sqrt{12n}}\right). \tag{18}$$

The region so defined will satisfy, approximately, condition (6). Denote by B the set of all regions $B(\theta_1)$ corresponding to all possible values of θ_1. We shall now consider whether or not the set B satisfies conditions (I) and (II). For this purpose, we fix an arbitrary sample point $(x_1, x_2, \cdots, x_n)$ and seek the set, say $S_B(x_1, x_2, \cdots, x_n)$ of those values of θ for which $(x_1, x_2, \cdots, x_n) \in B(\theta)$. Using definition (18) of the region $B(\theta)$, we find that, for $(x_1, x_2, \cdots, x_n) \in B(\theta)$, it is both necessary and sufficient that

$$\frac{\bar{x}}{\frac{1}{2} + \frac{\lambda(\alpha)}{\sqrt{12n}}} \leqq \theta \leqq \frac{\bar{x}}{\frac{1}{2} - \frac{\lambda(\alpha)}{\sqrt{12n}}}. \tag{19}$$

Just as in the case of set A, it is seen that the set $S_B(x_1, x_2, \cdots, x_n)$ covers a closed interval (19). It follows that B is a set of regions of acceptance and that the corresponding confidence limits are, say,

$$\begin{aligned} \underline{\vartheta}(X_1, X_2, \cdots, X_n) &= \frac{\overline{X}}{\frac{1}{2} + \frac{\lambda(\alpha)}{\sqrt{12n}}}, \\ \bar{\vartheta}(X_1, X_2, \cdots, X_n) &= \frac{\overline{X}}{\frac{1}{2} - \frac{\lambda(\alpha)}{\sqrt{12n}}}. \end{aligned} \tag{20}$$

The interval between these two limits is the confidence interval $\Delta(E)$ corresponding to the confidence coefficient α.

In order to bring out the delicate points of interpretation of formulae (14) and (20), we will use numerical examples. Thus, we shall select $\alpha = .95$ and substitute $n = 12$ [this particular value was selected in order to have less trouble with the square root of $12n$ in (20)]. Then formulae (14) and (20) reduce to

$$\underline{\theta} = X, \qquad \bar{\theta} = (1.284)X, \qquad \delta(E) = (0.284)X, \tag{21}$$

and

$$\underline{\vartheta} = (1.508)\overline{X}, \qquad \bar{\vartheta} = (2.970)\overline{X}, \qquad \Delta(E) = (1.462)\overline{X}, \tag{22}$$

respectively. In addition to the confidence limits, the lengths of the confidence intervals are also given.

The correct theoretical interpretations of formulae (21) and (22) are as follows.

Theoretical interpretation.—If the twelve observable random variables $X_1, X_2, \cdots, X_{12}$ *are completely independent and if each of them follows a uniform distribution between zero and* $\theta > 0$, *then, whatever the actual value* θ_1 *of* θ *may be, the probability that the greatest* X *of the* X_i *will not exceed* θ_1 *and, at the same time, that* $(1.284)X$ *will not be less than* θ_1 *is equal to the preassigned number* $\alpha = .95$,

$$P\{X \leqq \theta_1 \leqq (1.284)X \mid \theta = \theta_1\} = \alpha = .95.$$

Similarly and under the same conditions, we have

$$P\{(1.508)\overline{X} \leqq \theta_1 \leqq (2.970)\overline{X} \mid \theta = \theta_1\} = \alpha = .95.$$

From this probabilistic interpretation, we obtain the following operational interpretation.

Operational interpretation.—If the manner of obtaining the particular values of twelve variables $X_1, X_2, \cdots, X_{12}$ *is such that the assumption of their complete independence and uniform distribution between zero and some positive number* θ *is satisfied with a satisfactory approximation, then the long-run relative frequency of cases where* X *and* $(1.284)X$ *bracket* θ, *and also of those where* $(1.508)\overline{X}$ *and* $(2.970)\overline{X}$ *bracket* θ, *is approximately equal to* $\alpha = .95$.

Practical use of confidence intervals.—The above properties of confidence intervals were *deduced* from the specified assumptions regarding the observable random variables and, therefore, are the result of *deductive reasoning.* Having understood the meaning of these results, we may now *decide* (and this will be an act of will, not reasoning) to use these results in cases where it is desirable to have our actions adjusted to the value of θ which, unfortunately, is unknown. Our decision could be *to behave as if it were known for certain that the true value of* θ *lies between the lower and the upper confidence limits computed from actual observations.* The motivation behind this rule of behavior is simple: taking into account the operational interpretation of confidence intervals, we know that the long-run relative frequency of cases where our actions will be adjusted correctly, is equal to the number α which we have selected ourselves.

You will remember that this is just the requirement from a method of

estimation which a practical statistician may be reasonably expected to address to the theory.

I wish to emphasize the circumstance that the use of confidence intervals involves the following phases: (i) formulation of the problem, (ii) deductive reasoning leading to the solution of this problem and (iii) an act of will to adjust our behavior in accordance with the values of the confidence limits. In the past, claims have been made frequently that statistical estimation involves some mental processes described as *inductive reasoning*. The foregoing analysis tends to indicate that in the ordinary procedure of statistical estimation there is no phase corresponding to the description of "inductive reasoning." This applies equally to cases in which probabilities *a priori* are implied by the conditions of the problem and to cases in which they are not. In either case, all of the reasoning is deductive and leads to certain formulae and their properties. A new phase arrives when we decide to apply these formulae and to enjoy the consequences of their properties. This phase is marked by an act of will (not reasoning) and, therefore, if it is desired to use the adjective "inductive" in relation to methods of estimation, it should be used in connection with the noun "behavior" rather than "reasoning." The concept of "inductive behavior" is discussed in some detail in a book [6] in which it is treated as the motivational basis of the whole theory of statistics.

The operational interpretation of formulae (21) and (22) can be easily illustrated by a sampling experiment which you may wish to perform. In this it is convenient to use one of the published tables of random numbers.[7] As you know, ordinarily, tables of random numbers give groups of four digits, each digit selected at random from 0, 1, 2, $\cdots$, 9 with particular care that the consecutive selections be independent. Each such group can be considered as a decimal fraction with four digits. However, there is always the possibility of leaving out a digit or two or of adding a few more digits borrowed from the next group in the same line.

If we decide on a fixed number of digits, for example on three, then the consecutive groups in a column will produce an operational equivalent of

[6] J. Neyman: *First Course in Probability and Statistics*. Henry Holt and Co., New York, 1950, 350 pp.

[7] See, for example, the following tables:

(a) L. H. C. Tippett: *Random Sampling Numbers*. Tracts for Computers, No. xv, The University Press, Cambridge (Eng.), 1927, viii + 26 pp.

(b) M. G. Kendall and B. Babington Smith: *Tables of Random Sampling Numbers*. Tracts for Computers, No. xxiv, The University Press, Cambridge (Eng.), 1939, x + 60 pp.

(c) R. A. Fisher and Frank Yates: *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, London, 1938, 90 pp.

repeated independent observations of a random variable, say Y, which is discrete and is capable of assuming all values from .000 to .999, differing by .001. Moreover, all these particular values are approximately equiprobable. Obviously, Y, so defined, may be taken as an excellent approximation to the variable X distributed uniformly between zero and unity. Thus, if we select $\theta = 1$, the twelve independent "observations" of $X_1, X_2, \cdots, X_{12}$ can be read from any column of groups of three digits in a table of random numbers.

However, we need not limit ourselves to the value $\theta = 1$. In fact, you will find it instructive to select for your sampling experiment a set of, say, 100 different values (quite arbitrary) of θ. For example, one may be $\theta_1 = 1$, another $\theta_2 = .5$, still another $\theta_3 = 2$, etc. In order to obtain the simulated twelve observations of random variables uniformly distributed between zero and $\theta_k \neq 1$, it will be sufficient to take an appropriate number of digits in the table, write them as though they formed a decimal fraction and then multiply the result by θ_k. Naturally, if you want all the "measurements" of your "observable random variables" to be made with the same accuracy, you will have to use one more digit for $\theta_k = 10$ than for $\theta_k = 1$, etc.

Incidentally, I have just said that it would be instructive to embark on a sampling experiment with 100 arbitrarily selected θ's, all different. However, I am quite sure that after the fourth or fifth θ you will become convinced that the difference in the value of θ does not influence the relative frequency with which the confidence intervals cover the true θ and that, thereafter, you will use the simplest value of θ, namely, unity.

The sampling experiments are more easily performed than described in detail. Therefore, let us make a start with $\theta_1 = 1, \theta_2 = 2, \theta_3 = 3$ and $\theta_4 = 4$. We imagine that, perhaps within a week, a practical statistician is faced four times with the problem of estimating θ, each time from twelve observations, and that the true values of θ are as above although the statistician does not know this. We imagine further that the statistician is an elderly gentleman, greatly attached to the arithmetic mean and that he wishes to use formulae (22). However, the statistician has a young assistant who may have read (and understood) modern literature and prefers formulae (21). Thus, for each of the four instances, we shall give two confidence intervals for θ, one computed by the elderly Boss, the other by his young Assistant.

Using the first column on the first page of Tippett's tables of random numbers and performing the indicated multiplications, we obtain the following four sets of figures.

TABLE I

True θ	1st Sample $\theta_1 = 1$	2nd Sample $\theta_2 = 2$	3rd Sample $\theta_3 = 3$	4th Sample $\theta_4 = 4$
x_1	.295	1.368	2.334	.090
x_2	.416	.408	2.923	1.204
x_3	.273	1.113	2.826	2.996
x_4	.056	.902	.354	2.075
x_5	.275	.430	.212	.479
x_6	.587	1.383	1.905	1.563
x_7	.926	1.648	.424	.438
x_8	.200	1.583	2.112	.727
x_9	.956	1.877	.973	2.483
x_{10}	.824	.687	2.377	1.901
x_{11}	.566	1.819	2.631	.194
x_{12}	.101	1.845	.753	1.977
Arithmetic mean $\bar{x}$	.45625	1.25525	1.65200	1.34392
Greatest observation x	.956	1.877	2.923	2.996
Boss' conf. interval	$.688 \leqq \theta \leqq 1.355$	$1.892 \leqq \theta \leqq 3.728$	$2.490 \leqq \theta \leqq 4.907$	$2.026 \leqq \theta \leqq 3.992$
Asst.'s conf. interval	$.956 \leqq \theta \leqq 1.227$	$1.877 \leqq \theta \leqq 2.409$	$2.923 \leqq \theta \leqq 3.752$	$2.996 \leqq \theta \leqq 3.846$

The last two lines give the assertions regarding the true value of θ made by the Boss and by the Assistant, respectively. The purpose of the sampling experiment is to verify the theoretical result that the long run relative frequency of cases in which these assertions will be correct is, approximately, equal to $\alpha = .95$.

You will notice that in three out of the four cases considered, both assertions (the Boss' and the Assistant's) regarding the true value of θ are correct and that in the last case both assertions are wrong. In fact, in this last case the true θ is 4 while the Boss asserts that it is between 2.026 and 3.993 and the Assistant asserts that it is between 2.996 and 3.846. Although the probability of success in estimating θ has been fixed at $\alpha = .95$, the failure on the fourth trial need not discourage us. In reality, a set of four trials is plainly too short to serve for an estimate of a long run relative frequency. Furthermore, a simple calculation shows that the probability of at least one failure in the course of four independent trials is equal to .1855. Therefore, a group of four consecutive samples like the above, with at least one wrong estimate of θ, may be expected one time in six or even somewhat oftener. The situation is, more or less, similar to betting on a particular side of a die and seeing it win. However, if you continue the

sampling experiment and count the cases in which the assertion regarding the true value of θ, made by either method, is correct, you will find that the relative frequency of such cases converges gradually to its theoretical value, $\alpha = .95$.

Let us put this into more precise terms. Suppose you decide on a number N of samples which you will take and use for estimating the true value of θ. The true values of the parameter θ may be the same in all N cases or they may vary from one case to another. This is absolutely immaterial as far as the relative frequency of successes in estimation is concerned. In each case the probability that your assertion will be correct is exactly equal to $\alpha = .95$. Since the samples are taken in a manner insuring independence (this, of course, depends on the goodness of the table of random numbers used), the total number $Z(N)$ of successes in estimating θ is the familiar binomial variable with expectation equal to $N\alpha$ and with variance equal to $N\alpha(1 - \alpha)$. Thus, if $N = 100$, $\alpha = .95$, it is rather improbable that the relative frequency $Z(N)/N$ of successes in estimating θ will differ from α by more than

$$2\sqrt{\frac{\alpha(1 - \alpha)}{N}} = .042.$$

This is the exact meaning of the colloquial description that the long run relative frequency of successes in estimating θ is equal to the preassigned α.

Your knowledge of the theory of confidence intervals will not be influenced by the sampling experiment described, nor will the experiment *prove* anything. However, if you perform it, you will get an intuitive feeling of the machinery behind the method which is an excellent complement to the understanding of the theory. This is like learning to drive an automobile: gaining experience by actually driving a car compared with learning the theory by reading a book about driving.

Among other things, the sampling experiment will attract attention to the frequent difference in the precision of estimating θ by means of the two alternative confidence intervals (21) and (22). You will notice, in fact, that the confidence intervals based on X, the greatest observation in the sample, are frequently shorter than those based on the arithmetic mean $\overline{X}$. If we continue to discuss the sampling experiment in terms of cooperation between the eminent elderly statistician and his young assistant, we shall have occasion to visualize quite amusing scenes of indignation on the one hand and of despair before the impenetrable wall of stiffness of mind and routine of thought on the other.[8] For example, one can imagine the con-

[8] Sad as it is, your mind does become less flexible and less receptive to novel ideas as the years go by. The more mature members of the audience should not take offense. I, myself, am not young and have young assistants. Besides, unreasonable and stubborn

versation between the two men in connection with the first and third samples reproduced above. You will notice that in both cases the confidence interval of the Assistant is not only shorter than that of the Boss but is completely included in it. Thus, as a result of observing the first sample, the Assistant asserts that

$$.956 \leqq \theta \leqq 1.227.$$

On the other hand, the assertion of the Boss is far more conservative and admits the possibility that θ may be as small as .688 and as large as 1.355. And both assertions correspond to the same confidence coefficient, $\alpha = .95$! I can just see the face of my eminent colleague redden with indignation and hear the following colloquy.

Boss: "Now, how can this be true? I am to assert that θ is between .688 and 1.355 and you tell me that the probability of my being correct is .95. At the same time, you assert that θ is between .956 and 1.227 and claim the same probability of success in estimation. We both admit the possibility that θ may be some number *between* .688 and .956 or *between* 1.227 and 1.355. Thus, the probability of θ falling within these intervals is certainly greater than zero. In these circumstances, you have to be a nit-wit to believe that

$$\begin{aligned} P\{.688 \leqq \theta \leqq 1.355\} &= P\{.688 \leqq \theta < .956\} + P\{.956 \leqq \theta \leqq 1.227\} \\ &\quad + P\{1.227 < \theta \leqq 1.355\} \\ &= P\{.956 \leqq \theta \leqq 1.227\}." \end{aligned}$$

ASSISTANT: "But, Sir, the theory of confidence intervals does not assert anything about the probability that the unknown parameter θ will fall within any specified limits. What it does assert is that the probability of success in estimation using either of the two formulae (21) or (22) is equal to α."

Boss: "Stuff and nonsense! I use one of the blessed pair of formulae and come up with the assertion that $.688 \leqq \theta \leqq 1.355$. This assertion is a success only if θ falls within the limits indicated. Hence, the probability of success is equal to the probability of θ falling within these limits——."

ASSISTANT: "No, Sir, it is not. The probability you describe is the *a posteriori* probability regarding θ, while we are concerned with something else. Suppose that we continue with the sampling experiment until we have, say, $N = 100$ samples. You will see, Sir, that the relative frequency of successful estimations using formulae (21) will be about the same as that using formulae (22) and that both will be approximately equal to .95."

I do hope that the Assistant will not get fired. However, if he does, I would remind him of the glory of Giordano Bruno who was burned at the stake by the Holy Inquisition for believing in the Copernican theory of the solar system. Furthermore, I would advise him to have a talk with a physicist or a biologist or, maybe, with an engineer. They might fail to under-

---

individuals are found not only among the elderly but also frequently among young people.

stand the theory but, if he performs for them the sampling experiment described above, they are likely to be convinced and give him a new job. In due course, the eminent statistical Boss will die or retire and then——.

Now, let us forget the Boss and his Assistant and return to the important problem of the varying length of confidence intervals. By inspecting formulae (21) and (22), it is easy to see that for certain sample points the confidence interval (22), based on the arithmetic mean $\overline{X}$, is shorter than the corresponding confidence interval (21). When the greatest observation X is fixed, the arithmetic mean $\overline{X}$ of twelve positive observations may be arbitrarily close to $X/12$. Thus, the length of the corresponding confidence interval (22) may approach the lower bound of

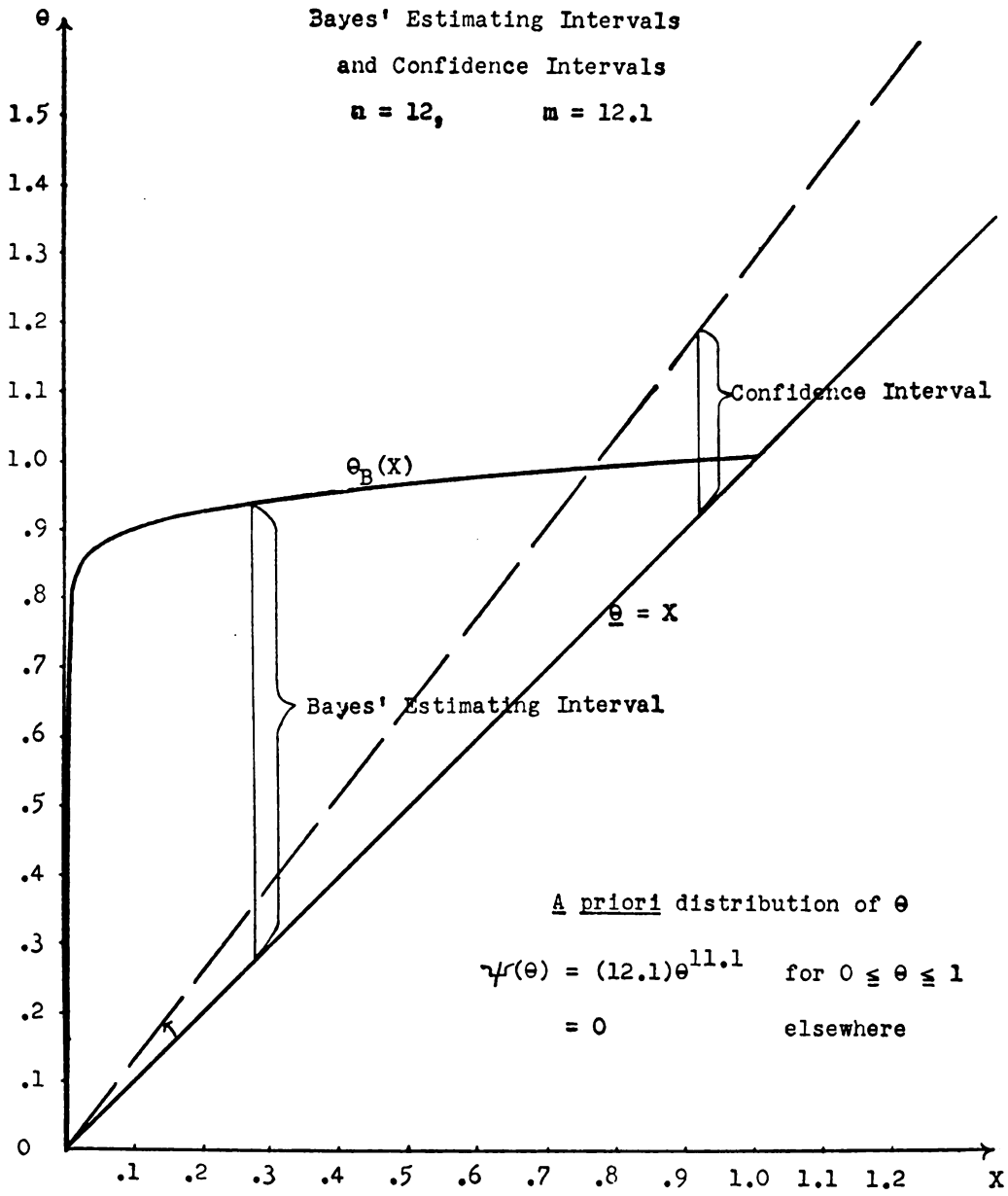
$$\frac{(1.462)X}{12} = (.122)X < \delta(E) = (.284)X.$$

This circumstance makes it intuitively clear how it happens that the use of either formulae (21) or (22) insures the same frequency of successes. However, if you perform the sampling experiment described above, you will notice that in the great majority of cases the confidence intervals computed from (21) are substantially shorter than those computed from (22). This empirical result suggests that, from the point of view of precision in estimating θ, formulae (21) are preferable to formulae (22). However, it is possible that some third pair of confidence limits can be invented, corresponding to the same confidence coefficient, which will give still better precision in estimating θ.

We are brought, thus, to the problem of a choice among the various possible confidence intervals and you will appreciate that this problem is of considerable practical importance and of great theoretical interest. Our first difficulty in attacking the problem consists in formulating it so that it has a definite mathematical meaning. In the case of known *a priori* distributions, the situation was simple because the Bayes' estimating interval corresponding to any given sample point is selected independently from those corresponding to other possible sample points. Therefore, we could simply seek that estimating interval which is shortest. With confidence intervals the situation is different because, instead of dealing directly with confidence intervals corresponding to particular sample points, we deal with regions of acceptance and the confidence interval corresponding to any given sample point depends upon the way in which the regions of acceptance are piled above this point. By shifting the regions of acceptance, it is possible to reduce to a minimum the length of the confidence interval corresponding to a specified sample point. However, it is intuitively clear that by so doing we shall increase the length of a great many other confidence intervals which correspond to different sample points. Thus, it appears that the

problem of "optimum" must concern not the length of particular confidence intervals taken separately, but the totality of these intervals.

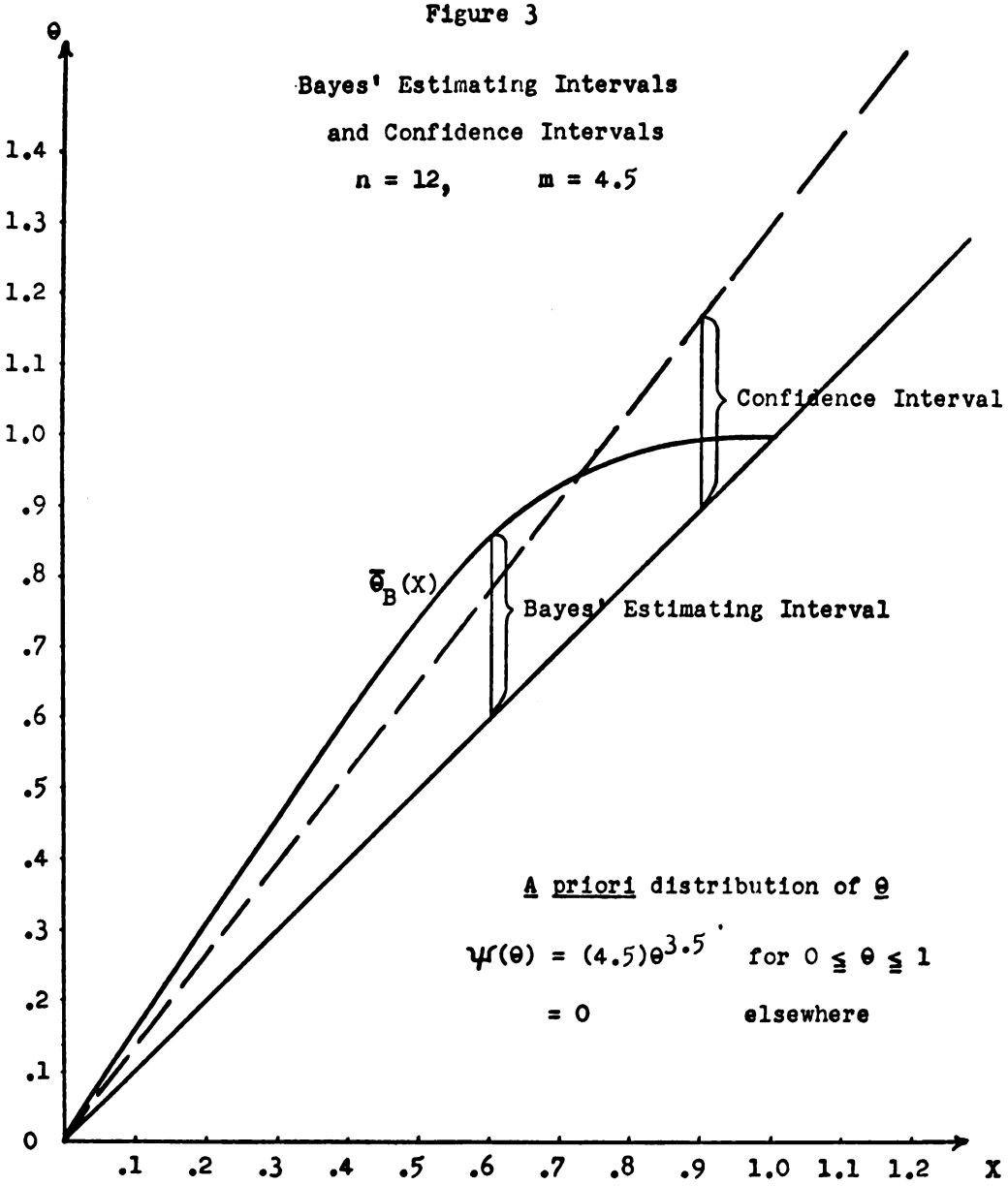
Figure 2



The problem I formulated in my paper of 1937 is based on the following considerations. The desirable property of a confidence interval is that it covers the true value of the estimated parameter θ with the preassigned frequency α. In so doing, the confidence interval also covers an infinity of

"false" values of θ. This, however, is a nuisance and should occur as rarely as possible. When one starts from this point of view, it is easy to give an exact definition of the *shortest confidence intervals.*

Figure 3



Definition.—The confidence interval $\delta(E)$ for estimating the parameter θ, corresponding to the confidence coefficient α, is called the shortest if, whatever be the alternative confidence interval $\Delta(E)$ corresponding to the same confidence coefficient α and whatever be two possible values θ_1 and θ_2 of the estimated parameter θ,

$$P\{\delta(E) \text{ C } \theta_1 \mid \theta_2\} \leqq P\{\Delta(E) \text{ C } \theta_1 \mid \theta_2\}.$$

Operationally, this means that, whatever be the true value θ_2 and whatever be the "false" value θ_1, this false value will be covered by $\delta(E)$ not more frequently than it will be covered by $\Delta(E)$.

It can be shown that the confidence interval (21) is the shortest in the sense of this definition. As you see, the "shortness" of the confidence interval $\delta(E)$ considered as a random interval used for purposes of estimation is consistent with the fact that for some particular sample points the length of interval (21) exceeds that of interval (22).

Unfortunately, in a great many cases of practical importance the shortest confidence intervals do not exist and we are forced to look for other possibilities. The study of these questions is an interesting and important part of the theory of estimation. However, its discussion would lead us far afield and all that I can do here is to refer you to my papers of 1937 and 1938 already quoted. At the present moment we will return to the problem of interpretation of confidence intervals and of Bayes' estimating intervals. Needless to say, a clear understanding of the difference between these two approaches to the problem is of fundamental and immediate importance to everyone concerned with estimation.

Figures 2 and 3 refer to the example of estimating θ, the only parameter involved in the probability density function (8) of $n = 12$ independent variables $X_1, X_2, \cdots, X_n$ all uniformly distributed between zero and $\theta > 0$. The two figures show confidence intervals (21). Thus the quantity measured on the axis of abscissae is X, the greatest of the twelve observations. The quantity measured on the axis of ordinates is θ. The heavy diagonal line has the equation $\underline{\theta}(X) = X$ and represents the lower confidence limit for θ. The dashed straight line above represents the upper confidence limit

$$\bar{\theta}(X) = (1.284)X.$$

Thus whatever be the observed value of X, the corresponding confidence interval for θ can be read directly from either of the two figures. The confidence intervals would be used when nothing is known about the *a priori* distribution of θ and the assertions regarding the true value of θ obtained from the graphs will be true with a long run relative frequency equal to .95.

Figures 2 and 3 also display the classical Bayes' estimating interval in addition to the confidence intervals. On both graphs the Bayes' estimating intervals correspond to an *a priori* distribution of θ of the same form,

$$\begin{aligned} \Psi(\theta) &= m\theta^{m-1} \qquad \text{for } 0 < \theta \leqq 1, \\ &= 0 \qquad\qquad \text{elsewhere.} \end{aligned} \tag{23}$$

However, in Figure 2 the value of m is $m = 12.1$ while in Figure 3, it is $m = 4.5$. In both cases, the lower end of the classical Bayes' estimating interval coincides with the lower confidence limit, $\underline{\theta}(X) = X$. On the other hand, the upper end of the classical Bayes' estimating interval is given by, say,

$$\theta_B(X) = [X^{m-n}(1 - \alpha) + \alpha]^{1/(m-n)},$$

where $\alpha = .95$ is the chosen confidence coefficient. The estimation of θ may consist in observing X and in asserting that θ lies within the corresponding Bayes' estimating interval. The long run relative frequency of successes will again be equal to α.

Upon inspecting the two figures, you are likely to have a feeling of surprise. Figure 2, especially, is striking because the Bayes' estimating intervals are so much wider than the corresponding confidence intervals over a very wide range of values of X. And yet, the Bayes' intervals are the shortest possible adjusted to the *a priori* distribution of θ of the particular kind indicated while the confidence intervals which assure the same long run frequency of successes in estimation will give this frequency quite irrespective of whether the *a priori* distribution is that visualized in Figure 2 or any other. Only if $X > .77$ (approximately) are the confidence intervals wider than the Bayes' intervals and then the difference in width is much milder than that, in favor of confidence intervals, for $X < .77$.

The answer is that, owing to the special form of the *a priori* distribution, larger values of θ will occur much more frequently than smaller ones and this is reflected also in the absolute distribution of X. This distribution is easily obtained as follows. We begin by writing the joint distribution of θ and X,

$$\begin{aligned} p_{\theta,X}(\theta, x) &= mn\theta^{m-n-1}x^{n-1} && \text{for } 0 < x \leqq 1 \text{ and } x \leqq \theta \leqq 1 \\ &= 0 && \text{elsewhere.} \end{aligned}$$

The absolute distribution of X is obtained by integrating

$$p_X(x) = \int_{-\infty}^{+\infty} p_{\theta,X}(\theta, x)\, d\theta = \frac{mn}{m - n}(x^{n-1} - x^{m-1}) \qquad \text{for } 0 < x \leq 1 \tag{24}$$

$$= 0 \text{ elsewhere.}$$

It is easy to verify that, if m and n exceed unity, this probability density vanishes at $x = 0$ and $x = 1$ and has its maximum at

$$x = \left(\frac{n - 1}{m - 1}\right)^{1/(m-n)}.$$

If $m = 12.1$ and $n = 12$, then this value exceeds .9. Thus in the situation represented in Figure 2, the most frequent values of X are close to unity and, consequently, the confidence intervals most frequently used will exceed the corresponding Bayes' shortest estimating intervals. A similar situation corresponds to Figure 3.

This reasoning explains only one aspect of the situation. To understand fully the other aspect we have to visualize the interpretation of the Bayes' estimating intervals in terms of frequencies. This may be done with reference to general human experience or, in order to speak in more concrete terms, with reference to a sampling experiment appropriately arranged so as to satisfy the hypotheses underlying Figure 2 and Figure 3.

It is essential that one makes the point clear that a sampling experiment which might illustrate all the properties of Bayes' estimating intervals is much more complicated than the one discussed above whose purpose was to illustrate the working of confidence intervals. In dealing with confidence intervals we were at liberty to select an arbitrary set of positive numbers and to consider these numbers as the true values of θ. Then it was an easy matter to use the tables of random numbers in order to obtain a sample of twelve observations following the distribution (8). Now we have to begin, as it were, earlier and create a machinery for obtaining a set of consecutive values of θ following the *a priori* distribution appropriate to Figure 2 and/or Figure 3. No arbitrary selection of the true θ's is allowed.

After this point has been settled in one way or another (those who are familiar with the arranging of sampling experiments will have no difficulty with this step), we proceed to obtain sampled values of $X_1, X_2, \cdots, X_{12}$, corresponding to each value of θ already determined. As in the case of confidence intervals (21), we shall be interested not in all twelve sample values but only in the greatest of them, X. The frequency distribution of this variable will correspond to the probability density (24). Now suppose that the first sample ascribes to X the value, $x = .500$. The Bayes' estimating interval corresponding to this value, as read from Figure 2, extends from .500 to .967, approximately. In order to interpret this interval (.500, .967) in detail, we would have to continue the sampling experiment for quite some time until we observed $x = .500$ another time, then still another time, etc. In short, for the interpretation of the Bayes' interval (.500, .967), we need a long sequence of outcomes of the sampling experiment in which the greatest of the twelve observations is equal to .500. It is obvious that the actual performance of this experiment is impractical unless it is performed using the most modern high speed computing machines. However, this should not preclude us from discussing it.

Imagine that, after we have repeated the sampling experiment a few million times, each time first determining a fresh value of θ in accordance

with probability density function (23) and then getting the twelve values of the X's, we have finally selected a set, say $S(.5)$, of some 100 cases in which the value of X was exactly equal to .500. This set, $S(.5)$, is basic for the interpretation of the Bayes' interval (.500, .967). Naturally, the values of θ corresponding to all experiments in $S(.5)$ will be different in general and we shall visualize the distribution of these values. In accordance with what was explained in yesterday's conference, this distribution will correspond approximately to the *a posteriori* probability density function,

$$p_{\theta \mid x} = \frac{m-n}{1-x^{m-n}}\theta^{m-n-1} \qquad \text{for } x \leqq \theta \leqq 1,$$

$$= 0 \qquad \text{elsewhere,}$$

with $x = .500$ and *the interval* (.500, .967) *will be found the shortest of all those which include 95 per cent of the values of* θ.

This is the precise interpretation of the Bayes' shortest estimating intervals. If you compare the foregoing with our previous discussion, you will see that *confidence intervals do not have the property just described.* In fact, if we take under consideration any particular confidence interval, e.g. the confidence interval (.500, .642) corresponding to the same value of $x = .500$, the relative frequency of experiments forming the set $S(.5)$ in which $.500 \leqq \theta \leqq .642$ will depend on the *a priori* distribution of θ and, in general, will not be equal to .95. On the other hand, whatever be the *a priori* distribution of θ, the assertion regarding the value of θ in any particular case, based on the confidence interval, has the probability equal to $\alpha = .95$ of being correct.

Before concluding, we shall make a very brief review of early papers of several authors in which one can discern the germs of the theory of confidence intervals.

The idea of estimation by confidence intervals and by confidence regions is very clearly and faultlessly stated in a few last sentences of a paper by Hotelling [9] published in 1931. However, the statement of this idea was not followed by an attempt to develop a systematic theory. The relevant passage is very brief and its brevity must have contributed to its being overlooked by many readers including myself. In order to give full credit to Hotelling, I wish to reproduce the passage verbatim.

To means of a single variate it is customary to attach a "probable error," with the assumption that the difference between the true and calculated values is almost certainly less than a certain multiple of the probable error. A more precise way to follow

[9] Harold Hotelling: "The generalization of Student's ratio." *Annals of Math. Stat.*, Vol. 2 (1931), pp. 360–378.

out this assumption would be to adopt some definite level of probability, say $P = .05$, of a greater discrepancy, and to determine from a table of Student's distribution the corresponding value of t, which will depend on n; adding and subtracting the product of this value of t by the estimated standard error would give upper and lower limits between which the true values may with the given degree of confidence be said to lie. With T an exactly analogous procedure may be followed, resulting in the determination of an ellipse or ellipsoid centered at the point $\xi_1, \xi_2, \cdots, \xi_p$. Confidence corresponding to the adopted probability P may then be placed in the proposition that the set of true values is represented by a point within this boundary. (Harold Hotelling: *Annals of Math. Stat.,* Vol. 2, pp. 377–378.)

Next in turn, in the reverse chronological order, it would be necessary to refer to papers by R. A. Fisher concerned with the so-called "fiducial argument." The early papers of Fisher given to this subject definitely suggest the idea of confidence intervals. Later on, however, there appeared to be a substantial difference between the two theories. The relevant literature is extensively discussed in the next part of the present chapter.

Before either Hotelling or Fisher, the idea of confidence intervals is found in papers by E. B. Wilson [10] and Stanislas Millot.[11] Both authors are concerned with estimating the probability p of success postulated to be constant in n completely independent trials in which the success occurred exactly X times. Working independently, the two authors used similar arguments to deduce the approximate confidence interval for p, based on the assumption that the distribution of the standardized binomial variable, say

$$\frac{X - np}{\sqrt{np(1 - p)}},$$

is approximately normal. However, the conceptual background of the two papers is essentially different from the statement of the problem of confidence intervals and is limited to the view that it is reasonable to use the formula deduced. Wilson explains this clearly in his more recent paper [12] given to the same problem. In addition, the paper of Millot involves obvious misunderstandings which it may be useful to discuss. For this purpose, we shall deduce the formulae for the (approximate) lower and upper confidence limits for p.

We begin by postulating that $Y = X/n$ is a normal variable with expectation p and variance $p(1 - p)/n$, where p stands for the unknown true

[10] E. B. Wilson: "Probable inference, the law of succession, and statistical inference." *Jr. Amer. Stat. Assoc.,* Vol. 22 (1927), pp. 209–212.

[11] Stanislas Millot: "Sur la probabilité *a posteriori.*" *Comptes Rendus,* Paris Academy, t. 176 (1923), pp. 30–32.

[12] E. B. Wilson: "On confidence intervals." *Proc. Nat. Acad. Sc.,* Vol. 28 (1942), pp. 88–93.

probability of success and may be any number $0 < p < 1$. Following the steps indicated in the earlier part of this conference we proceed to construct a system A of regions of acceptance, say $A(p^*)$. Obviously, the sample space of the variable Y is limited to the interval $0 \leqq Y \leqq 1$. Fix any possible value p' of p, and determine in W a region $A(p')$ satisfying condition (6). Considerations of simplicity suggest that the region $A(p')$ be represented by an interval, say from $a(p')$ to $b(p')$ with

$$0 \leqq a(p') \leqq b(p') \leqq 1.$$

Then condition (6) implies that, if p' happens to be the true value of p,

$$P\{a(p') \leqq Y \leqq b(p') \mid p = p'\} = \alpha,$$

where α is the adopted confidence coefficient. Using the postulate that Y is a normal variable, we can rewrite this condition as

$$\frac{\sqrt{n}}{\sqrt{2\pi p'(1-p')}} \int_{a(p')}^{b(p')} e^{-n(y-p')^2/2p'(1-p')}\, dy = \alpha. \tag{25}$$

Obviously, equation (25) does not determine $a(p')$ and $b(p')$ uniquely. In fact, it is possible to select $a(p')$ arbitrarily, provided the value selected is not too large, and then to determine $b(p')$ to satisfy condition (25). Considerations of simplicity suggest that $a(p')$ and $b(p')$ be symmetrically placed about p' so that

$$a(p') = p' - \Delta,$$
$$b(p') = p' + \Delta.$$

Now, if λ satisfies the condition

$$\frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{+\lambda} e^{-x^2/2}\, dx = \alpha,$$

then

$$\Delta = \lambda \sqrt{\frac{p'(1-p')}{n}}$$

and the region of acceptance $A(p')$ is defined by the double relation,

$$p' - \lambda \sqrt{\frac{p'(1-p')}{n}} \leqq Y \leqq p' + \lambda \sqrt{\frac{p'(1-p')}{n}}.$$

We shall adopt this definition of $A(p')$ for every possible value p' of p and test whether or not the set A of all such regions satisfies conditions (I) and (II). For this purpose we fix an arbitrary possible value y of Y and look for the values p for which $y \in A(p)$. The search reduces to the solution

with respect to p' of the two inequalities defining region $A(p')$. If we drop the primes and perform easy algebra, we find

$$|Y - p| \leqq \lambda \sqrt{\frac{p(1-p)}{n}},$$

$$Y^2 - 2pY + p^2 \leqq \lambda^2 \frac{p - p^2}{n},$$

or

$$p^2\left(1 + \frac{\lambda^2}{n}\right) - 2p\left(Y + \frac{\lambda^2}{2n}\right) + Y^2 \leqq 0. \tag{26}$$

Since the coefficient of p^2 is positive, the left hand side of inequality (26) is negative only if the two roots of the quadratic are real and the value of p is contained between the smaller and the larger of these roots. Denoting the roots by $p_1(Y)$ and $p_2(Y)$, we have

$$p_1(Y) = \frac{Y + \frac{\lambda^2}{2n} - \frac{\lambda}{\sqrt{n}}\sqrt{Y(1-Y) + \frac{\lambda^2}{4n}}}{1 + \frac{\lambda^2}{n}},$$

$$p_2(Y) = \frac{Y + \frac{\lambda^2}{2n} + \frac{\lambda}{\sqrt{n}}\sqrt{Y(1-Y) + \frac{\lambda^2}{4n}}}{1 + \frac{\lambda^2}{n}}$$

and it is seen that they are always real. Thus, the set of values of p for which $Y \epsilon A(p)$ extends over the closed interval,

$$p_1(Y) \leqq p \leqq p_2(Y).$$

Consequently, the regions $A(p)$ are regions of acceptance and $p_1(Y)$ and $p_2(Y)$ are a pair of confidence [13] limits for p, corresponding to the confidence

[13] Incidentally, a closer analysis shows that these limits possess the defect of being "biased." While covering the "true" value with the prescribed relative frequency α, the confidence interval $[p_1(Y), p_2(Y)]$ covers certain "false" values of p even more frequently. This fact is due to the adopted symmetry of regions of acceptance. By dropping the requirement of symmetry, it is possible to obtain somewhat "shorter" confidence intervals corresponding to the same confidence coefficient and covering the false values of p less frequently than the true value. However, this advantage of the unbiased confidence intervals is, in this case, not very important. When n is large, then

coefficient α. Thus, if we use the formulae for $p_1(Y)$ and $p_2(Y)$ to make assertions regarding the true value of p in the form $p_1(Y) \leqq p \leqq p_2(Y)$, the probability of this assertion being true is (approximately) equal to α.

In connection with confidence intervals for the binomial p, I should bring to your attention the fact that Clopper and Pearson [14] have produced convenient graphs from which these intervals can be read directly. I should also like to note that, if one does not assume n large enough for the validity of the normal approximation, then one has to deal with an extended notion of confidence intervals in which the probability of the true value of the parameter being covered is *at least equal to* (instead of equal or approximately equal to) the chosen confidence coefficient. The method of constructing such intervals is discussed and illustrated in the joint publication [15] of Matuszewski, Supinska and myself.

As mentioned, the formulae for $p_1(Y)$ and $p_2(Y)$ were deduced both by E. B. Wilson and by Stanislas Millot. However, Millot interprets them as a result relating to the probability *a posteriori*, with which, in reality, these formulae have no connection whatever. Moreover, the following passage translated from Millot's note indicates that his idea regarding the operational properties of the interval $[p_1(Y), p_2(Y)]$ were in disaccord with the basic concepts of confidence intervals.

Millot writes:

> It is useful to record the results of the various experiments made, because, frequently but to a variable extent, the study of partial series of such experiments may allow us to reduce the uncertainty regarding the true value of the probability p. To each partial series, as well as to the total series of observations there corresponds an interval for p, the boundaries of which are determined from formulae (5). Evidentally, the probability p is contained in the common part of all the intervals thus obtained.

The statement "Evidentally, the probability p is contained in the common part of all the intervals thus obtained," has no probabilistic meaning and, therefore, has no room within the theory of confidence intervals. If we admit the possibility of a *lapsus linguae* and try to reword this statement in conformity with the concepts of confidence intervals, we would obtain something like this: "The probability that the common part of all intervals thus obtained will bracket the true value of p is even greater than the

---

the difference between the unbiased intervals and the ones deduced here is insignificant. On the other hand, when n is small, then the normal approximation which we used here is inadequate.

[14] C. J. Clopper and E. S. Pearson: "The use of confidence or fiducial limits illustrated in the case of the binomial." *Biometrika*, Vol. 26 (1934), pp. 405–413.

[15] T. Matuszewski, J. Neyman and J. Supinska: "Statistical studies in questions of Bacteriology. Part I. The accuracy of the 'Dilution Method.'" *Supplement, Jr. Roy. Stat. Soc.*, Vol. 2 (1935), pp. 63–82.

confidence coefficient α." However, even this interpretation does not bring the idea of Millot within the framework of confidence intervals. In the latter, the probability of bracketing the true value of the parameter applies to a completely specified rule, i.e., to a pair of functions, such as $p_1(Y)$ and $p_2(Y)$, defined for all possible values of the observable random variables. As I have shown this morning, this probability coincides with the probability of the sample point falling within the region of acceptance corresponding to the true value of the parameter estimated. Also, I have shown that the regions of acceptance are uniquely determined by the confidence limits. Now, while implying what should be our assertion regarding p when the several confidence intervals overlap, Millot does not say a word about this assertion when the confidence intervals fail to overlap. Thus, Millot's estimating intervals are not defined for all combinations of values of the observable random variables and, therefore, the regions of acceptance are not defined. As a result, without further specification of the estimation procedure contemplated, it is impossible to assert anything about the probability that it will lead to a correct assertion.

In addition to the note discussed, Millot has published a few more notes in the same volume of *Comptes Rendus*. However, the general idea behind these notes diverges more and more from the basic concept of confidence intervals expressed at the beginning of the first note.

If we go further back, we can trace the idea of confidence intervals, very vaguely expressed, in the writings of "Student." Also, although no explicit statement has been found, it is possible that the idea of confidence intervals may have been behind the publications of Markoff and even of Gauss, concerned with what is now called "best unbiased estimates."

This brings us to the question of how the use of confidence intervals can be considered a justification for the use of best unbiased estimates and of maximum likelihood estimates (when they are consistent and efficient). Actually, the argument is in favor of a broader category of estimates, having the property that they are asymptotically normal about the true value of the parameter with minimum asymptotic variance. This point of view was brought out in my paper of 1934 already quoted.

Let θ be a parameter to be estimated using a large number n of observable random variables, the totality of which will be denoted by a single letter X_n. Let, further, $F_n(X_n)$ denote a function of X_n and $\sigma_n(\theta)$ a function of n and θ but not of X_n, having the property that, as $n \to \infty$, for all $\lambda > 0$,

$$\lim_{n \to \infty} P \left\{ \frac{| F_n(X_n) - \theta |}{\sigma_n(\theta)} \leqq \lambda \right\} = \frac{2}{\sqrt{2\pi}} \int_0^\lambda e^{-x^2/2}\, dx.$$

By selecting an appropriate λ, the integral in the right hand side may be made equal to the chosen confidence coefficient α. Hence, if n is sufficiently

large, the probability in the left hand side of this equation will differ but very little from α. But this probability coincides with the probability,

$$P\{F_n(X_n) - \lambda\sigma_n(\theta) \leqq \theta \leqq F_n(X_n) + \lambda\sigma_n(\theta)\},$$

which indicates that the two limits, say

$$\underline{\theta} = F_n(X_n) - \lambda\sigma_n(\theta) \quad \text{and} \quad \bar{\theta} = F_n(X_n) + \lambda\sigma_n(\theta),$$

have the properties of an (approximate) confidence interval corresponding to the confidence coefficient α. It is true that, without the knowledge of θ, these limits may be impossible to compute. The important fact, however, is that the length of the interval indicated is $2\lambda\sigma_n(\theta)$ and thus is a fixed multiple of $\sigma_n(\theta)$. Thus, if a number of functions like $F_n(X_n)$ are available for estimating θ, the tendency towards the greatest precision in estimation as measured by the length of the approximate confidence interval implies a preference for those estimates for which the asymptotic variance $\sigma_n^2(\theta)$ is smallest. This, then, is one of the rational justifications, which may be brought forward, for the use of best unbiased and maximum likelihood estimates in the (frequent) cases in which they are asymptotically normal and efficient. It will be noticed, however, that this justification has nothing to do with any sort of principle or axiom but is based on purely utilitarian considerations of consequences of repeated application of the procedure described.

It may be worthwhile to emphasize that the justification for the use of best unbiased estimates explicitly stated by Gauss is a different one. As Laplace had already noticed, the process of estimating an unknown parameter θ may be compared with a game of chance in which a statistician, using an estimate $F_n(X_n)$, may lose a positive quantity [when $F_n(X_n) \neq \theta$] or may break even [when $F_n(X_n) = \theta$], but in which he can never gain. The quantity lost is, therefore, a monotone increasing function of the absolute value of the difference $|F_n(X_n) - \theta|$, the nature of which, however, cannot be deduced from the general circumstances of the problem of estimation. Thus, this loss function, say $L[F_n(X_n) - \theta]$, may be selected arbitrarily in conformity with each particular problem of estimation. Once the loss function is selected, the goodness of any particular estimate $F_n(X_n)$ may be measured by the expectation, say "risk,"

$$R[F_n(X_n), \theta] = E\{L[F_n(X_n) - \theta]\},$$

of the loss which will be incurred when $F_n(X_n)$ is used as an estimate of θ.

Laplace himself studied certain problems on the assumption that the loss due to an error in estimation is directly proportional to the absolute value of the error. On the other hand, Gauss noticed that various results

became more elegant if the loss is assumed to be proportional to the square of the error committed so that

$$L[F_n(X_n) - \theta] = [F_n(X_n) - \theta]^2.$$

Upon reflecting on the general nature of errors of measurements, in particular, on the possibility of systematic errors, Gauss found it necessary to impose on the estimate $F_n(X_n)$ another condition, that of unbiasedness, expressed by identity,

$$E[F_n(X_n)] \equiv \theta.$$

It will be seen that the two conditions, one of the unbiasedness of $F_n(X_n)$ and the other of minimum expected loss measured by the square of the error, formulate the now familiar problem of best unbiased estimates. All this was reported to the Königliche Societät der Wissenschaften in Göttingen on February 15, 1821, and subsequently published in Latin. A German translation by A. Börsch and P. Simon appeared in a book under the general title, *Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrick Gauss,* Berlin, 1887, pp. v + 208. I enter into these bibliographical details partly in an attempt to correct a confusion to which I unwittingly contributed by attributing to Markoff the basic theorem on least squares. See, for example, F. N. David and J. Neyman: "An extension of the Markoff theorem on least squares," *Stat. Research Memoirs,* Vol. II (1938), pp. 105–116. As R. L. Plackett pointed out in his "A historical note on the method of least squares" (*Biometrika,* Vol. 36 (1949), pp. 458–460), the theorem that I ascribed to Markoff was discovered by Gauss and published in the remarkable memoir just quoted.[16]

Early in the present century, the idea of the loss function attracted the attention of F. Y. Edgeworth who, under the label of "detriment" discussed it in a number of his papers, published in *Mind* and in the *Journal of the Royal Statistical Society.* In one of these papers (*Jr. Roy. Stat. Soc.,* Vol. 71, 1908, and 72, 1909) he was in search of the "most advantageous" estimates, that is, such estimates as would, in large samples, minimize the average detriment. Edgeworth, anticipating Fisher by thirteen years, conceived the conviction that the "most advantageous" estimates (in present day terminology, asymptotically normal estimates with minimum asymptotic variance) are those obtained by the "genuine inverse method," or as we say now, following Fisher, the maximum likelihood estimates.

After Edgeworth, the idea of the loss function was lost from sight for more than two decades, to be revived in a paper by E. S. Pearson and

[16] For more recent developments in this direction, see E. W. Barankin and John Gurland: "On asymptotically normal, efficient estimators: I," *University of California Publications in Statistics,* Vol. 1, No. 6 (1951), pp. 89–130.

myself, "The testing of statistical hypotheses in relation to probabilities *a priori*" (*Camb. Phil. Soc. Proc.,* Vol. 29 (1933), pp. 492–510). Unfortunately, at that time we were not aware of the fact that the idea was not new. Also, being preoccupied with other problems, we just mentioned the idea as a possible approach to the problem without attempting to do anything concrete. The real revival of the idea of the loss function and of the associated risk function began with the entry on the scene of statistical research of Abraham Wald. Combining the concept of loss with another concept of minimax (also outlined in the above publication of 1933), Wald has initiated a new branch of statistical theory and, followed by Wolfowitz and a host of younger searchers, brought it to a remarkable level of elegance and generality. The principal results obtained in this direction are summarized in the recent book of Wald: *Statistical Decision Function* (Wiley, New York, 1950, pp. ix + 179).

Part 3. Fiducial Argument and the Theory of Confidence Intervals

(This section has been reproduced from *Biometrika,* Vol. 32 (1941), pp. 128–150, through the courtesy of the Editor, Professor E. S. Pearson.)

1. INTRODUCTION

The theory of confidence intervals was started by the present author about 1930. At that time it was taught in lectures given both at the University and at the Central College of Agriculture, Warsaw, Poland. The theory found immediate practical applications, and before any theoretical paper was published, a booklet (Pytkowski, 1932) [1] appeared giving numerical confidence intervals for means and for regression coefficients. The term "confidence interval" is a translation of the original Polish "przedział ufności." The author's theoretical results appeared two years later (Neyman, 1934). At almost the same time the first tables and graphs of confidence intervals were published (Clopper & Pearson, 1934) in a paper which gave a remarkably clear explanation of the difference between the new approach to the problem of estimation and the old one, by means of Bayes's theorem.

The first publication on fiducial argument (Fisher, 1930) anticipated the booklet of Pytkowski by two years. The present author overlooked this article for some time. However, when preparing his paper of 1934, he was already acquainted with it and also with the next paper (Fisher, 1933) on a similar subject. Although Fisher's method of approach was entirely different from the author's, the numerical identity of Fisher's fiducial limits

[1] The references cited are given in full at the end of this Part, on pages 253–254.

with the confidence limits in the author's theory, and also some of Fisher's early comments, suggested to the author that the two theories are essentially the same. Accordingly, and owing to the difference in dates of publications, the author considered his own work as an extension of the previous results of Fisher. This was clearly stated in the author's paper of 1934.

Apart from the above points of agreement the author had found certain passages and conceptions in the publications of Fisher which were difficult for him to understand and to reconcile with what was essential in the theory of confidence intervals. They included "fiducial probability" and "fiducial distribution of a parameter." However, the author was inclined to think that these were, more or less, *lapsus linguae,* difficult to avoid in the early stages of a new theory. This attitude was clearly expressed in the paper of 1934. That paper was read before a meeting of the Royal Statistical Society and was followed by a public discussion recorded in the Society's *Journal.* Fisher took part in the discussion, and it was a great surprise to the author to find that, far from recognizing them as misunderstandings, he considered fiducial probability and fiducial distributions as absolutely essential parts of his theory. As a result, the author began to doubt whether the two theories were, in fact, equivalent. These doubts were only increased by Fisher's insistence that the calculation of fiducial distributions and fiducial limits must be limited to cases where sufficient statistics exist (Fisher, 1936), and by his warnings against inconsistencies in the theory of confidence intervals.

When questioned on the subject, the author could not conceal his doubts and they were published (Neyman, 1938*a*). Subsequent publications by other authors appear to be divided. Some, e.g. the very important papers by Wald (1939) and by Wald & Wolfowitz (1939), deal with the theory of confidence intervals, entirely ignoring fiducial theory. Others (Starkey, 1938; Sukhatme, 1938; Yates, 1939), at the other extreme, work on the ground of fiducial argument and ignore the confidence intervals. There is also an intermediate group of authors with an almost continuous spectrum of opinions. Pitman (1939), in a very interesting paper on estimation of location and scale parameters, states that the two theories "are essentially the same and that their two points of view are both necessary for a full comprehension of the theory of estimation." And a few pages further: "I at first called it the fiducial probability function, but finally decided to shorten the name by dropping the word 'probability.' "

Next we find the statement (Bartlett, 1939) that "by a distribution of fiducial type we shall mean a distribution providing at least confidence intervals in the sense of Neyman." This statement is used in an argument (Bartlett, 1936, 1939) that, as a distribution deduced by Fisher (1936) does not seem to provide confidence limits, there must be some error in the

deduction. A similar point of view, but with a stronger leaning towards confidence intervals, is expressed by Welch (1939). In this paper various general claims of Fisher are analyzed, essentially from the point of view of confidence intervals, and tested on appropriate examples. Among other things it is found that the fears of inconsistencies in the theory of confidence intervals are unfounded.

A quite different school of thought is represented by Jeffreys (1940), according to which the fiducial approach to the problem of estimation is completely equivalent with that by inverse probability.

Fisher (1937, 1939*a*, 1939*b*) and Yates (1939) emphatically deny that there is an error in Fisher's paper of 1936. On the contrary, it is said that the results then published were obscured by the controversy arising from Bartlett's confusion about the nature of fiducial argument. Also, especially in earlier papers (1930, 1933, 1936), Fisher is equally emphatic on the distinction between the fiducial and the inverse probability approaches to the problem of estimation.

The above survey shows that there is an interesting divergence of opinions as to what is essential in the fiducial theory in general and as to whether it is in any way connected with the theory of confidence intervals. The perusal of all the literature quoted does not allow the present author to form any precise opinion as to the first of these questions. On the other hand, there now seems to be sufficient ground for answering the second, concerning the relationship between the two theories. The purpose of the present paper is to show that there is none. The relevant points concerning this question, which were possible to establish on the ground of earlier literature, are explained in excellent papers by Pearson (1939) and Welch (1939), with the final conclusion that, in spite of various differences, the two theories are closely related. However, fresh evidence provided by papers of Fisher (1939*a*, 1939*b*) and Yates (1939) shows that no such relation exists and that the authors suspecting it were misled by the incompleteness of earlier writings concerning fiducial argument.

As a result of the present paper it may be found expedient, for the sake of clarity, to avoid confusion of terminologies appropriate to the two theories. Instead of writing, as some authors do, on "fiducial *or* confidence" limits, it may be preferable to discuss "fiducial limits" or "confidence limits," as the case may be, separately.

2. BASIC IDEAS IN THE THEORY OF CONFIDENCE INTERVALS

The key to understanding the theory of confidence intervals is in being clear about what might be called the classical point of view in the theory of probability. This theory was originally built up to answer questions

about *how frequently* a given combination of throws will occur in a long series of games of dice. Thus, the probability of a certain combination found to be, say, 1/5, implies that this combination would appear in about 20% of a long series of actual games. This agreement may, but need not, be observed. In the latter case, we would say that the assumptions underlying the deduction were not realized by the actual experiments. The dice used were perhaps "biased," and so forth. The point is that, whenever it is said that a given set of probabilities does refer to some phenomena, then it is understood that the relative frequencies of various aspects of the phenomena, in a long series of trials, are approximately equal to corresponding probabilities. This is just what the author calls the classical point of view in the theory of probability. It is excellently explained by v. Mises (1939), but is more general than the definition of probability adopted by that author.[2]

Apart from the classical point of view on probability, there is another. It considers the probabilities as measures of rational belief in the truth of a given proposition. Here the agreement between the probability and some relative frequency is not essential.

The theory of confidence intervals was built up to give a solution of problems of estimation which would have a clear frequency interpretation, characteristic of the classical point of view. Consider a set E of n observable random variables, $x_1, \cdots, x_n$, and assume as given that the function $p(E \mid \theta_1, \theta_2, \cdots, \theta_s)$ represents its elementary probability law. Here $\theta_1, \cdots, \theta_s$ represent certain parameters whose values are unknown.

The above should be interpreted as follows. There are some actual trials T which are able to determine the values of the x's. There are also some numbers $\vartheta_1, \vartheta_2, \cdots, \vartheta_s$, unknown to us, such that, whatever be a region w in the space of the x's, the integral of $p(E \mid \vartheta_1, \vartheta_2, \cdots, \vartheta_s)$ taken over this region is approximately equal to the relative frequency with which the point E, as determined by the trials T, falls within that region w. The problem of estimating one of the parameters, e.g. θ_1, consists in using just one system of the x's as determined by the trials T to calculate ϑ_1 approximately. Alternatively, it may consist in calculating an interval $(a, a + d)$ which "presumably" covers ϑ_1.

The original approach to this problem is based on Bayes's theorem. Denote by $p(\theta_1, \theta_2, \cdots, \theta_s)$ the elementary probability law of the θ's. Then

$$p(\theta_1, \theta_2, \cdots, \theta_s \mid E') = \frac{p(\theta_1, \cdots, \theta_s)p(E' \mid \theta_1, \cdots, \theta_s)}{\int \cdots \int p(\theta_1, \cdots, \theta_s)p(E' \mid \theta_1, \cdots, \theta_s)d\theta_1 \cdots d\theta_s} \quad (1)$$

[2] It will be noticed that the classical point of view on probability does not imply any particular definition of that concept. It is not suggested that the one adopted by v. Mises is the only one that could be consistently used.

will be the relative probability law, or the probability law *a posteriori* of all the θ's given the observed system E' of the values of the x's. It can be used to calculate the most probable value of θ_1. Alternatively, given a number $d > 0$, the law can be used to find the interval $(a, a + d)$ such that the *a posteriori* probability

$$P\{a + d > \theta_1 > a \mid E'\}$$

is greatest.

Our attitude towards this kind of solution, dictated by the classical point of view on probability, depends on circumstances and may be twofold.

The circumstances of the problem may imply not only that the x's but also that the θ's are random variables and that the function $p(\theta_1, \cdots, \theta_s)$ could be used to calculate the relative frequencies of various combinations of values of the θ's. Such situations are rare, but they do occasionally occur, especially in problems of genetics and of mass production. If the function $p(\theta_1, \cdots, \theta_s)$ is implied by the problem considered, then the probability $P\{a + d > \theta_1 > a \mid E'\}$ has a clear frequency interpretation, as follows. Imagine a long sequence, S, of cases where the θ's vary according to the above law and the x's are determined by the particular trials considered. Pick from this sequence S a subsequence $S(E')$ of such trials in which the experiments determined the same system of values of the x's, namely, the system E'. Naturally, the value of θ_1 in cases belonging to $S(E')$ would vary. But, if the functions $p(E \mid \theta_1, \cdots, \theta_s)$ and $p(\theta_1, \cdots, \theta_s)$ do have the presumed relation to the trials considered, it will be found that among all the intervals of length d, the interval $(a, a + d)$ will contain the value of θ_1 more frequently than any other, and that this frequency will be approximately equal to $P\{a + d > \theta_1 > a \mid E'\}$. It follows that, if the function $p(\theta_1, \cdots, \theta_s)$ is implied by the circumstances of the problem of estimation, the use of the formula (1) is perfectly legitimate from the point of view of the classical theory of probability.

The situation is quite different when the circumstances of the problem do not imply the *a priori* probability law. This is most frequently the case. Moreover, usually there are serious difficulties in considering the θ's as random variables. Jeffreys (1939) advises the use of formula (1) also in such cases, with a function $p(\theta_1, \cdots, \theta_s)$ invented for the purpose. He claims that the conclusions drawn in this way are valid, provided that the function used is just the one that he suggests. The present author would not question this statement on condition that the word "valid," or any other such description, is not given any significance beyond that described above. In other words, there seems to be no reason why we should not agree to *call* the above conclusions "valid in the sense of Jeffreys." On the other hand, it seems essential to be clear that any probability calculated from (1), with any function $p(\theta_1, \cdots, \theta_s)$ not implied by the actual prob-

lem, need not and, generally, will not have any relation to relative frequencies. It will not be the probability in the classical sense of the word and, therefore, persons who would like to deal only with classical probabilities, having their counterparts in the really observable frequencies, are forced to look for a solution of the problem of estimation other than by means of the theorem of Bayes.

This solution (Neyman, 1937, 1938*b*) may be obtained as follows. Consider the case where the circumstances imply that the x's, forming a system E, are random variables with the probability law $p(E \mid \theta_1, \theta_2, \cdots, \theta_s)$, where $\theta_1, \theta_2, \cdots, \theta_s$ are unknown. Denote by $\underline{\theta}(E)$ and $\bar{\theta}(E)$ two functions of the x's. Obviously, if E is random then these functions will also be random variables.

DEFINITION 1. *If the functions* $\underline{\theta}(E)$ and $\bar{\theta}(E)$ *possess the property that, whatever be the possible value* ϑ_1 *of* θ_1 *and whatever be the values of the unknown parameters* $\theta_2, \theta_3, \cdots, \theta_s$, *the probability*

$$P\{\underline{\theta}(E) \leq \vartheta_1 \leq \bar{\theta}(E) \mid \vartheta_1, \theta_2, \cdots, \theta_s\} \equiv \alpha, \tag{2}$$

then we will say that the functions $\underline{\theta}(E)$ *and* $\bar{\theta}(E)$ *are the lower and the upper confidence limits of* θ_1, *corresponding to the confidence coefficient* α. *The interval* $[\underline{\theta}(E), \bar{\theta}(E)]$ *is called the confidence interval for* θ_1.

In spite of the complete simplicity of the above definition, certain persons have difficulties in following it. These difficulties seem to be due to what Karl Pearson (1938) used to call routine of thought. In the present case the routine was established by a century and a half of continuous work with Bayes's theorem. It may be useful, therefore, to give a few illustrations.

Assume that $s = 2$, that θ_1 may have only the five values 1, 2, 3, 4, and 5, and that, at the same time, θ_2 may vary continuously between zero and 1. To satisfy Definition 1, the only requirement on the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ is that

$$P\{\underline{\theta}(E) \leq \vartheta \leq \bar{\theta}(E) \mid \vartheta, \theta_2\} \equiv \alpha \tag{3}$$

for all values of $\vartheta = 1, 2, 3, 4$, and 5, and for θ_2 varying between (0, 1). The probabilities (2) and (3) are, therefore, *not* the probabilities of θ_1 falling within any limits. On the contrary, they are the probabilities of the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ falling on both sides of a specified number ϑ. These probabilities are to be calculated from the given function $p(E \mid \theta_1, \theta_2)$ with the value of θ_1 set equal to the same number ϑ. The result must be totally independent of the values of $\theta_2, \cdots, \theta_s$ and must equal α.

It is known (Neyman, 1935*b*; Feller, 1938) that in certain cases no such functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ exist. Then there are ways of modifying the formulation of the problem, for example, requiring that the probability on the left of (2) be at least equal to α, and so forth. In other cases, there will be an

infinity of pairs of confidence limits all corresponding to the same α. In this case, the practical statistician is at liberty to choose among them.

Let us now consider the frequency interpretation of the solution of the problem of estimation by means of confidence intervals. Suppose that some two functions $\underline{\theta}(E) \leq \bar{\theta}(E)$ possess property (2) with some large value of α, say $\alpha = 0.99$. Their use in practice would consist of (i) observing the values E' of the x's, (ii) calculating the corresponding values of the confidence limits $\underline{\theta}(E')$ and $\bar{\theta}(E')$, and (iii) *stating* that the true value ϑ_1 of θ_1 lies between $\underline{\theta}(E')$ and $\bar{\theta}(E')$. The justification is simple and perfectly in line with the classical point of view of probability: in the course of many applications, the relative frequency of cases in which the statement $\underline{\theta}(E) \leq \vartheta_1 \leq \bar{\theta}(E)$ is correct will be approximately equal to $\alpha = 0.99$, whether or not the parameters for estimation are the same in all cases.

The word "stating" above is put in italics to emphasize that it is not suggested that we can "conclude" that $\underline{\theta}(E') \leq \vartheta_1 \leq \bar{\theta}(E')$, nor that we should "believe" that ϑ_1 is actually between $\underline{\theta}(E)$ and $\bar{\theta}(E)$. In the author's opinion, the word "conclude" has been wrongly used in that part of statistical literature dealing with what has been termed "inductive reasoning." Moreover, the expression "inductive reasoning" itself seems to involve a contradictory adjective. The word "reasoning" generally seems to denote the mental process leading to knowledge. As such, it can only be deductive. Therefore, the description "inductive" seems to exclude both the "reasoning" and also its final step, the "conclusion." If we wish to use the word "inductive" to describe the results of statistical inquiries, then we should apply it to "behaviour" and not to "reasoning." The fact that a given pair of functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ satisfies the identity (2) may be "deduced" from the properties of the function $p(E \mid \theta_1, \cdots, \theta_s)$. Earlier trials may show characteristics in the empirical distribution of the x's which seem in agreement with the function $p(E \mid \theta_1, \cdots, \theta_s)$. On these grounds, after observing the values of the x's in a case where the θ's are unknown and calculating $\underline{\theta}(E')$ and $\bar{\theta}(E')$, we may *decide* to behave as if we actually knew that the true value ϑ_1 of θ_1 were between $\underline{\theta}(E')$ and $\bar{\theta}(E')$. This is done as a result of our *decision* and has nothing to do with "reasoning" or "conclusion." The reasoning ended when the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ were calculated. The above process is also devoid of any "belief" concerning the value ϑ_1 of θ_1. Occasionally we do not behave in accordance with our beliefs. Such, for example, is the case when we take out an accident insurance policy while preparing for a vacation trip. In doing so, we surely act against our firm belief that there will be no accident; otherwise, we would probably stay at home. This is an example of inductive behaviour.

Obviously, if there are many different pairs of functions, $\underline{\theta}(E)$ and $\bar{\theta}(E)$, all corresponding to the same α, our choice of the one to use must be based on the

detailed study of their properties. For example, if it appears that the difference between one pair, $\bar{\theta}_1(E) - \underline{\theta}_1(E)$, is always (or most frequently) smaller than that between some other pair, then we would probably prefer to use the first. The problem of determining the confidence limits and of studying their properties forms the subject of the theory of confidence intervals.

3. NECESSARY AND SUFFICIENT CONDITIONS FOR A PAIR OF FUNCTIONS TO BE CONFIDENCE LIMITS

Let $a(E) \leq b(E)$ be any two single-valued functions of the x's determined for all possible systems of their values. Denote by W the space of the x's and by ϑ_1 one of the possible values of θ_1. Finally, let $A(\vartheta_1)$ denote the region in the space W composed of all points E which satisfy the double inequality,

$$a(E) \leq \vartheta_1 \leq b(E). \tag{4}$$

It was proved (Neyman, 1937) that for the two functions, $a(E)$ and $b(E)$, to be the lower and upper confidence limits for the parameter θ_1, it is necessary and sufficient that, whatever be the possible value ϑ_1 of θ_1, the probability

$$P\{E \in A(\vartheta_1) \mid \theta_1 = \vartheta_1\} \equiv \alpha. \tag{5}$$

The identity refers to the arbitrary variation of $\theta_2, \cdots, \theta_s$.

This condition will be used below to show that a certain pair of functions does not represent the confidence limits. For this purpose, the following steps will be taken: We shall select a convenient value ϑ_1 of the estimated parameter θ_1 and determine the region $A(\vartheta_1)$ as in (4). Next, we shall substitute this same value ϑ_1 instead of the parameter θ_1 in the elementary probability law of the variables considered, getting $p(E \mid \vartheta_1, \cdots, \theta_s)$. This last function will be integrated over $A(\vartheta_1)$ to find the probability $P\{E \in A \mid \theta_1 = \vartheta_1\}$ as in the left-hand side of (5). But this integral will be dependent on the values of the other parameters involved, showing that the identity (5) is not satisfied. The conclusion will be that the particular functions considered are not confidence limits.

4. DIFFERENCES BETWEEN THE THEORY OF CONFIDENCE INTERVALS AND THE THEORY OF FIDUCIAL ARGUMENT

In this section we will consider examples treated both from the point of view of confidence intervals and of fiducial argument. These will be selected to illustrate both the conceptual and the numerical differences between the two theories.

(i) *Evidence of conceptual differences between the two theories.*—The first results obtained concerning confidence intervals (Neyman, 1934) refer to the case where all the n observable variables x_i are mutually independent, nor-

mally distributed, have the same though unknown standard error σ, and expectations $\mathcal{E}(x_i)$ which are linearly connected with some $s < n$ unknown parameters $p_1, p_2, \cdots, p_s$, so that

$$\mathcal{E}(x_i) = a_{i1}p_1 + a_{i2}p_2 + \cdots + a_{is}p_s. \tag{6}$$

Here the a's are supposed to be known and to form a non-singular matrix. Denote by θ any linear combination of the same p's, that is

$$\theta = b_1p_1 + b_2p_2 + \cdots + b_sp_s, \tag{7}$$

with known b's not all equal to zero. In these circumstances, a confidence interval for θ is given by

$$F - St_\alpha \leq \theta \leq F + St_\alpha, \tag{8}$$

where F denotes the best unbiased estimate of θ (David & Neyman, 1938), S the estimate of the standard error of F, and t_α the value of the "Student"-Fisher t corresponding to the number of degrees of freedom $n - s$ and to $P = 1 - \alpha$. The application of more recent theory (Neyman, 1935*b*) shows that the confidence intervals (8) have distinct advantages over any others by satisfying the definition (Neyman, 1937) of the "short unbiased system of type B_1." Without entering into these details, we shall consider the particular case where $s = 1$, $a_{i1} = 1$ and $b_1 = 1$. This will be the case if all the x's come from the same unknown normal population and it is desired to estimate its mean, $\theta = \mathcal{E}(x_i)$. In that case $F = \bar{x}$ and

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}. \tag{9}$$

As mentioned, the general confidence interval (8) was discussed in lectures about 1930, and in 1932 a publication appeared using the concept and the formula (8).

As far as is known, the first full discussion of the corresponding result in the fiducial theory was given by Fisher a few years later (Fisher, 1935, 1936), and here is the relevant passage from the second paper.

> If a sample of n observations, $x_1, \cdots, x_n$, has been drawn from a normal population having a mean value μ, and if from the sample we calculate the two statistics $\bar{x} = \Sigma x_i/n$ and $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$, $\cdots$, "Student" has shown (1925) [3] that the quantity t, defined by the equation
>
> $$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}, \tag{10}$$
>
> is distributed in different samples in a distribution dependent only from the size of the sample, n. It is possible, therefore, to calculate, for each value of n, what value of t will be

[3] Actually, of course, this result appeared earlier ("Student," 1908).

exceeded with any assigned frequency, P, such as 1% or 5%. These values of t are, in fact, available in existing tables (Fisher, 1925–34).

It must now be noticed that t is a continuous function of the unknown parameter, the mean, together with observable values, $\bar{x}$, s and n, only. Consequently the inequality $t > t_1$ is equivalent to the inequality

$$\mu < \bar{x} - \frac{st_1}{\sqrt{n}}, \tag{11}$$

so that this last inequality must be satisfied with the same probability as the first. This probability is known for all values of t_1, and decreases continuously as t_1 is increased. Since, therefore, the right-hand side of the inequality takes, by varying t_1, all real values, we may state the probability that μ is less than any assigned value, or the probability that it lies between any assigned values, or, in short, its probability distribution, in the light of the sample observed.

It is of some importance to distinguish such probability statements about the value of μ, from those that would be derived by the method of inverse probability, from any postulated knowledge of the distribution of μ in the different populations which might have been sampled. . . . To distinguish it from any of the inverse probability distributions derivable from the same data it has been termed the *fiducial* probability distribution, and the probability statements which it embraces are termed statements of fiducial probability.

In the next section we shall analyze the above passage in detail and show exactly where and how it conflicts with the classical theory of probability and thus with the theory of confidence intervals. Here we will mention only that it is ambiguous. Just this kind of ambiguity, which is also found in the earlier papers (Fisher, 1930, 1933), is probably responsible for a number of authors, including the present one, thinking that the fiducial theory and the theory of confidence intervals are linked.

In a few years it was found necessary to reinterpret formula (11). This was done by Fisher himself (1939*b*) and, somewhat more clearly but on the same lines, by Yates (1939). It will be seen from the following quotation from Yates's paper that the above passage by Fisher certainly does not contain everything which is *now* considered essential in the fiducial theory and that the presumption of any link between the latter and the theory of confidence intervals is unfounded. Yates's more relevant sentences are italicized by the present author.

While explaining the meaning of the fiducial distribution of the mean μ of a normal population, Yates mentions that the fiducial distribution of σ^2 is given by

$$\frac{1}{\sigma^2} = \frac{\chi_0^2}{\Sigma(x_i - \bar{x})^2}, \tag{12}$$

where χ^2 has its usual distribution with $n - 1$ degrees of freedom.

It can then be shown that, for a value of μ equal to μ_ϵ and a given s, the value of $\bar{x}$ in subsequent samples would be as small as that observed in a fraction ϵ of the samples, *provided that the actual distribution of σ^2 is the same as the fiducial distribution given above.*

In this form, however, the statement is open to objection on the ground that in subsequent samples σ may in fact be distributed in any manner, and that s will certainly vary from sample to sample. To avoid this objection we must frankly recognize *that we have here introduced a new concept into our methods of inductive inference, which cannot be deduced by the rules of logic from already accepted methods.* . . . That is . . . the form of fiducial statement which is implicit in the t test as ordinarily used by practical experimenters. . . . It must be recognized as essentially different from the statement that t will exceed t_ϵ in a fraction ϵ of all experiments. The latter is true for any given fixed σ or any set of σ's. The former (i.e., the fiducial statement, J.N.) *is true for a given s when σ is taken to be fiducially distributed in the appropriate distribution.* . . . The logical difference between the two approaches (fiducial and inverse probability, J.N.) should, however, be recognized. The approach by inverse probability enables fiducial statements about μ to be derived from the classical theory of probability, without the introduction of any new principle, but only at the cost of postulating a particular *a priori* distribution of σ. *In the fiducial approach such a priori postulation is regarded as inadmissible, but in order to discard it a new principle, that of utilizing the fiducial distribution of σ, must be introduced.* . . . Once the principle is accepted it is possible, given $\bar{x}$ and s, to make formal and exact statements of the fiducial type about μ which are independent of all prior knowledge of σ. *If the principle is not accepted, then it appears that we must either assume an a priori distribution of σ, or deny that there is any possibility of making fiducial statements about μ.*

The present author is unable to understand the exact meaning of what is called "fiducial statements about μ." However, his conclusion is that their conceptual nature must be quite different from that dealt with in the theory of confidence intervals. This conclusion is based on the fact that all the difficulties described by Yates as inherent in the fiducial theory are non-existent in the theory of confidence intervals. Applications of the latter require no new principle "which cannot be deduced by the rules of logic," no assumption that this or that unknown parameter follows any specified distribution, and have no connexion with Bayes's theorem. To make the situation absolutely clear, imagine a sequence of normal populations $\pi_1, \pi_2, \cdots, \pi_m, \cdots$, with their means $\theta_1, \theta_2, \cdots, \theta_m, \cdots$ and their standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_m, \cdots$. Imagine that out of each population π_m we have a random sample Σ_m of n individuals, with its mean $\bar{x}_m$ and an estimate of the corresponding variance S_m^2 as in (9). The theory of confidence intervals guarantees that the relative frequency with which $\bar{x}_m - t_\alpha S_m$ will fall short of the corresponding θ_m and, at the same time $\bar{x}_m + t_\alpha S_m$ will exceed this same number θ_m, will be, within an error of sampling, equal to α. An incredulous reader may easily check this by a sampling experiment. In this he will be at liberty to keep θ_m and/or σ_m constant, or to vary them at his pleasure, without any restriction. Of course, the distributions of the populations sampled should be more or less normal and the sampling should be random. It follows from the above passages of Yates that if the requirements above are satisfied but no new principles accepted, then we have to deny that there is any possibility of making fiducial statements about θ_m. If so, then the nature of the latter is

different from those involved in the application of the theory of confidence intervals.

The comparison of the above comments by Yates with those of Fisher gives a curious impression. Where Yates sees so many difficulties and restrictions, Fisher mentions none. Yet this very publication of Yates is fully endorsed by Fisher (1939*b*).

(ii) *Numerical differences between the two theories.*—Besides establishing the existence of conceptual differences, it is essential to show that the two theories may give different numerical results. We may conclude from the discussion above that the application of confidence intervals requires fewer restrictions. But there is a logical possibility that, when both theories are applicable, they give the same numerical result. The following example shows that this is not the case and that fiducial limits need not satisfy the definition of confidence limits.

The example that we are going to discuss refers to the problem of estimating the difference, say δ, between the means of two populations of which it is known only that both are normal. Denote by

$$\left.\begin{array}{llll} x_{1,1}, & x_{1,2}, \cdots, & x_{1,n}, \\ x_{2,1}, & x_{2,2}, \cdots, & x_{2,n'}, \end{array}\right\} \tag{13}$$

two random samples to be drawn from these populations and let $n \leq n'$. The confidence limits for δ have been very elegantly obtained by Bartlett. He did not publish his results himself but they are briefly mentioned in a paper by Welch (1938). The tendency towards a greater generality of presentation resulted in certain complications. The following is a less general but simplified statement of the results.[4] Assume that the x's in (13) are numbered in the order in which they will be given by observation. Otherwise, randomize the second series. Next calculate n differences

$$u_i = x_{1,i} - x_{2,i}, \qquad (i = 1, 2, \cdots, n). \tag{14}$$

If $\mathcal{E}(x_{1,i}) = \theta + \delta$ and $\mathcal{E}(x_{2,i}) = \theta$, then $\mathcal{E}(u_i) = \delta$. If the S.D.'s of the two populations sampled are σ and σ', then the S.E. of u_i will be $(\sigma^2 + \sigma'^2)^{\frac{1}{2}}$. The consecutive u's will be normal and independent and the problem of estimating the difference between the means of two normal populations will be reduced to that of estimating the mean of one population of the u's. Its solution is given by the confidence interval

$$\bar{u} - St_{(\alpha)} \leq \delta \leq \bar{u} + St_{(\alpha)}, \tag{15}$$

where S has an obvious meaning and $t_{(\alpha)}$ is to be taken with $n - 1$ degrees of freedom.

[4] Apart from these, the same author has obtained certain relevant results referring to the case where $n = n' = 2$ (Bartlett, 1936).

Again, an experiment consisting in repeated sampling of pairs of normal populations will show that, whatever be θ, δ, σ, σ', whether constant or varying in an absolutely arbitrary manner, the relative frequency of cases in which the statement about δ in the form of (15) will be true will be approximately equal to α. The above solution of the problem, elegant as it is, is only a partial one. The results of Bartlett do not tell us whether the family of systems of confidence intervals found by him exhausts all the possibilities and whether it is possible to construct intervals which would be, in one sense or another, shorter than those given by (15). These are interesting and important problems and we may hope to have them solved.

Remark added in 1951: Since the above lines were first published in *Biometrika,* it became apparent that, in the ideas described, Bartlett was anticipated by V. Romanovsky (*Atti del Congresso Internazionale dei Matematici,* Vol. 6, 1928, pp. 103–105). Also, the problem of an optimum solution within the category outlined was solved by Henry Scheffé (*Annals of Math. Stat.,* Vol. 14, 1943, pp. 35–44). Later on, Scheffé's solution was extended to an analogous but somewhat more complicated problem by E. W. Barankin (*Proc., First Berkeley Symposium on Math. Stat. and Probability,* 1945/46, pp. 433–449).

A result in fiducial theory corresponding to, but not equivalent with, formula (15) has been published by Fisher (1936):

Let us suppose that a sample of n observations has yielded a mean, $\bar{x}$, and an estimated variance of the mean, s^2, so that $s^2 = \Sigma(x_i - \bar{x})^2/n(n-1)$; then we know that if μ is the mean of the population

$$\mu = \bar{x} + st, \tag{16}$$

where t is distributed in "Student's" distribution. Similarly, for the mean of a second population, of which we have n' observations, we may write

$$\mu' = \bar{x}' + s't', \tag{17}$$

where t' is distributed in "Student's" distribution with $n' - 1$ degrees of freedom, independently of t. If now

$$\mu' - \mu = \delta, \qquad \bar{x}' - \bar{x} = d, \tag{18}$$

we find that

$$\epsilon = \delta - d = s't' - st, \tag{19}$$

and since s' and s are known, the quantity represented on the right has a known distribution, though not one which has been fully tabulated. The equation may be written

$$\epsilon = \sqrt{(s^2 + s'^2)}(t' \cos R - t \sin R), \tag{20}$$

where $\tan R = s/s'$, so that R is a known angle. If t and t' be taken as the co-ordinates of a point on a plane, the frequency of the observations falling within any area of the plane is calculable. The points for which θ has any given value lie on a straight line, at a distance from the origin $\pm\epsilon/(s^2 + s'^2)^{\frac{1}{2}}$, and making an angle R with the axis of t. The fiducial probability that ϵ exceeds any given value is the frequency in the area above this line. If n and n' are both increased, the distribution of ϵ tends to be normal and independent of R; when R is 0° or 90° the distribution is of "Student's" form. In general it involves n, n', and R and for any chosen probability, therefore, requires a table of triple entry.

As the reader will notice, no restrictions are mentioned and it is not suggested that for the practical application of the results any assumption is needed concerning the variability of the variances of the populations sampled. Neither is there any suggestion of any new principle that may be involved. We will return to this point below.

Following the publication of Fisher just quoted, and on his advice, Sukhatme published a table (Sukhatme, 1938). The quantity tabled may be denoted by $f(n, n', R)$ and represents the root of the equation

$$\int_{-\infty}^{+\infty} \left\{ G(t) \int_{\kappa}^{+\infty} H(t')\, dt' \right\} dt = 0.025, \tag{21}$$

where $G(t)$ and $H(t')$ are "Student's" distributions with $n - 1$ and $n' - 1$ degrees of freedom respectively, while

$$\kappa = \frac{f(n, n', R)}{(s^2 + s'^2)^{\frac{1}{2}} \cos R} + t \tan R. \tag{22}$$

It follows from the context that $f(n, n', R)$ so calculated is the value such that the fiducial probability of its being exceeded by $|\epsilon|/(s^2 + s'^2)^{\frac{1}{2}}$ is equal to 0.05. In other words, the values $f(n, n', R)$ are the fiducial 5% limits of $|\epsilon|/(s^2 + s'^2)^{\frac{1}{2}}$. As $\epsilon = \delta - d$, if the presumption that the fiducial limits necessarily lead to confidence intervals be true then this means that the double inequality

$$\bar{x}' - \bar{x} - f(n, n', R)\sqrt{s^2 + s'^2} \leq \delta \leq \bar{x}' - \bar{x} + f(n, n', R)\sqrt{s^2 + s'^2} \tag{23}$$

must be the confidence intervals for $\delta = \mu' - \mu$. But it is easy to see that the functions on the extreme parts of (23) do not satisfy the conditions, explained in § 3 above, necessary and sufficient for them to be the confidence limits. Take $\delta = 0$ and denote simply by A the region in the space of the x's including all the points in which the inequality (23) is satisfied. Take the probability law of the x's and put $\delta = 0$ in it, that is, $\mu' = \mu$. It will be seen that the integral $I(A)$ of this probability law taken over A depends on the ratio $\rho = \sigma/\sigma'$ of the two σ's appropriate to the two populations sampled and, thus, that it does not satisfy the identity (5).

Condition (23) defining the region A does not involve the particular x's but only the means $\bar{x}$, $\bar{x}'$, and the variances s^2 and s'^2. Consequently, to calculate $I(A)$ we may start with the probability law of those four variables

$$p(\bar{x}, \bar{x}', s, s') = \frac{c}{\sigma^n \sigma'^{n'}} s^{n-2} s'^{n'-2} \times \exp\left\{ -\frac{n(\bar{x} - \mu)^2}{2\sigma^2} - \frac{n'(\bar{x}' - \mu)^2}{2\sigma'^2} - \frac{n(n-1)s^2}{2\sigma^2} - \frac{n'(n'-1)s'^2}{2\sigma'^2} \right\}, \tag{24}$$

where c is a purely numerical constant and does not involve any of the parameters. This function must be integrated over the region A defined by (23) or by the equivalent inequality

$$\frac{|\bar{x}' - \bar{x}|}{\sqrt{s^2 + s'^2}} \leq f(n, n', R). \tag{25}$$

In dealing with it, we have to remember that R is not a constant but is connected with s and s' by the equation $\tan R = s/s'$. The required integral, or probability, of $\bar{x}$, $\bar{x}'$, s, and s' satisfying (25) will be more easily calculated if we introduce a new system of variables, u, v, R, and s_0. These will be connected to the old system as follows:

$$\left.\begin{aligned} \bar{x} &= \mu + us_0 \sin R, \\ \bar{x}' &= \mu + vs_0 \cos R, \\ s &= s_0 \sin R, \\ s' &= s_0 \cos R. \end{aligned}\right\} \tag{26}$$

The Jacobian J of the transformation is easily found to be

$$J = s_0^3 \sin R \cos R. \tag{27}$$

The limits of variation of the new variables are as follows:

$$\left.\begin{aligned} &-\infty < u, v < +\infty, \\ &0 \leq s_0, \\ &0 \leq R \leq \tfrac{1}{2}\pi. \end{aligned}\right\} \tag{28}$$

The probability law of the new variables will be

$$p(u, v, s_0, R) = \frac{c}{\sigma^n \sigma'^{n'}} s_0^{n+n'-1} e^{-\psi^2 s_0^2/2} \sin^{n-1} R \cos^{n'-1} R, \tag{29}$$

with

$$\psi^2 = \frac{nu^2 \sin^2 R}{\sigma^2} + \frac{n'v^2 \cos^2 R}{\sigma'^2} + \frac{n(n-1) \sin^2 R}{\sigma^2} + \frac{n'(n'-1) \cos^2 R}{\sigma'^2}. \tag{30}$$

Inequality (25) will be equivalent to

$$|v \cos R - u \sin R| \leq f(n, n', R). \tag{31}$$

As this does not involve s_0 the integration with respect to this variable can be carried out within the extreme limits of its variation. As a result further integrations may be performed on the probability law of u, v, R,

$$p(u, v, R) = \int_0^\infty p(u, v, s_0, R)\, ds_0$$

$$= \frac{c}{\sigma^n \sigma'^{n'}} \frac{\sin^{n-1} R \cos^{n'-1} R}{\psi^{n+n'}}, \tag{32}$$

where c is again a numerical constant.

Further integration may be conveniently carried out as follows. Substitute a new variable z for the variable v so that

$$v = \frac{z + u \sin R}{\cos R}, \qquad \frac{\partial v}{\partial z} = \frac{1}{\cos R}. \tag{33}$$

Keep z constant within the limits $|z| \leq f(n, n', R)$ prescribed by (31) and integrate for u from $-\infty$ to $+\infty$. The result is

$$p(z, R) = \frac{c \sin^{n-2} R \cos^{n'-2} R}{\sigma^{n-1} \sigma'^{n'-1} \sqrt{n\sigma'^2 + n'\sigma^2}}$$

$$\times \left\{ \frac{nn'}{n\sigma'^2 + n'\sigma^2} z^2 + \frac{n(n-1)}{\sigma^2} \sin^2 R + \frac{n'(n'-1)}{\sigma'^2} \cos^2 R \right\}^{-(n+n'-1)/2} \tag{34}$$

The integration is completed by an easy substitution for z

$$I(A) = c\rho^{n'-1} \int_0^{\frac{1}{2}\pi} \left\{ \frac{\sin^{n-2} R \cos^{n'-2} R}{\{n(n-1) \sin^2 R + n'(n'-1)\rho^2 \cos^2 R\}^{(n+n'-2)/2}} \right.$$

$$\left. \times \int_0^{wf} \frac{dz}{(1+z^2)^{(n+n'-1)/2}} \right\} dR, \tag{35}$$

with $f = f(n, n', R)$ and

$$w^2 = \frac{\dfrac{nn'}{n\sigma'^2 + n'\sigma^2}}{\dfrac{n(n-1)}{\sigma^2} \sin^2 R + \dfrac{n'(n'-1)}{\sigma'^2} \cos^2 R}. \tag{36}$$

By inspecting (35) it is more or less evident that $I(A)$ must depend on the value of ρ. However, to avoid any doubt in this respect, it was thought useful to calculate $I(A)$ for a few values of ρ. This was done by Miss Elizabeth Scott of the Statistical Laboratory, University of California, and it is a pleasure to record the author's indebtedness to her. The calculations involved supplementing the tables of Sukhatme for a denser set of values of R. The calculated values of $I(A)$ are:

$n = 12, \quad n' = 6$

ρ	$I(A)$
0.1	0.966
1.0	0.960
10.0	0.934

Thus the functions representing the fiducial limits for δ do not satisfy the conditions necessary and sufficient for them to be the confidence limits of the parameter in question. It follows that if pairs of normal populations forming a long sequence are sampled and the extreme parts of the double inequality (23) calculated, then the relative frequency of cases where the prediction of the value of δ by means of these inequalities will be correct need not be equal to the expected 0.95. It will depend on the value of ρ and, if this is uncertain, this frequency will be unknown. Subsequent comments by Fisher (Fisher, 1939*a*) seem to indicate that the frequency in question is expected to approach 0.95 only if the ratio ρ is not constant but follows a certain fiducial distribution. It is noteworthy that no such restriction is to be found in the original work quoted above. On the other hand, it is more or less in line with those restrictions formulated by Yates.

5. VIEWS OF M. S. BARTLETT AND R. A. FISHER

The controversy in which the main contributors are Bartlett (Bartlett, 1936, 1939) and Fisher (Fisher, 1937, 1939*a*, 1939*b*) seems to be based on a misunderstanding. Presuming that the fiducial limits are always equal to confidence limits, Bartlett was puzzled by Fisher's results concerning δ just quoted, and suspected an error. The subsequent elaborations by Fisher and Yates amount to a confirmation that the values of $f(n, n', R)$ as tabled by Sukhatme do not provide the confidence intervals. But both authors are emphatic that there is no error in the original deductions, and that Bartlett misunderstood the problem. It is unthinkable that these four unanimous papers are mistaken and, therefore, we must accept the conclusion that the presumption of intrinsic identity between fiducial and confidence limits is unfounded.

But it must be pointed out that, before the appeal to extra-logical principles was published, there was much to be said in favor of the opinion that the solution of Fisher, as quoted above, and the work of Sukhatme both involved errors in the algebra of probability laws. It also seems that, apart from establishing that the fiducial theory and the theory of confidence

intervals are distinct, it will be of some interest to analyze Fisher's work in detail and to point out exactly where and how it diverges from the rules of ordinary theory of probability on which the theory of confidence intervals is based.

When a system of observable phenomena is treated mathematically, it is essential to be clear on exactly what is assumed as given or as known. For example, when trying to calculate the area of land from a certain set of measurements, it is essential to be clear as to assumptions made concerning the shape of the land considered. The available data may be consistent with a number of such assumptions, e.g. that the surface considered is a plane or that it is spherical with a given radius, etc. Whichever of these hypotheses is accepted as given, the applications of the appropriate formulae will give mutually consistent results. But they would not generally be consistent if one part of the calculations were made on one hypothesis and another on a contradictory one. The differences may be small, but in mathematics there are really no "small" nor "large" inconsistencies. There are simply inconsistencies. Needless to say, the choice of exactly what is to be accepted as given must be made to attain the greatest conformity with empirical facts. But this is a question which need not be discussed here.

The above general principle also applies to the applications of probability. There we must be clear as to exactly what are the phenomena or the variables which we agree to consider as *random* in a given inquiry. In practice, of course, the random variable will be the one whose value at the moment is uncertain and is being determined "by chance." If X is considered as a random variable, the premises of the mathematical problem must include some assumptions as to the relative frequencies with which X assumes its possible values. These assumptions may vary in specificity, but they must be present in the premises.

Any number or variable which is not random must be clearly recognized as such. For some time such non-random numbers were called constants. This was more or less satisfactory with constant *numbers*. But Fréchet (Fréchet, 1937) has noticed that we may also consider *variables* which are not random and has invented useful terms to describe them. These are "nombre certain," "fonction certaine," etc. We will translate these terms by "sure number" and "sure function." The thousandth digit in the expansion $\pi = 3.1415\ldots$ is a sure number, although totally unknown to me. Denote by $f(n)$ the relative frequency of 0's among the first n digits of the same expansion of π. This will be a sure function. On the other hand, if $\phi(n)$ denotes the number of errors that may be made when calculating π to n places of decimals, then $\phi(n)$ *may be* considered as a random function of n. Considerations of this kind would imply those of a considerable sequence S of similar attempts to calculate π, by the same person or by

different persons of a specified category, in which the values of $\phi(n)$ will vary, as we shall say, at random. It is with respect to just such a sequence of determinations of the values of the function $\phi(n)$ that our probability statements will refer. For example, if we either start or finish our calculations with the probability equal to 0.25 of $\phi(n)$ being between any two sure numbers a, b, then the applicational statement is that about 25% of the numbers of the sequence S satisfy the inequality $a < \phi(n) < b$.

It is important to notice that the sequence S *may* consist of just one member; then all the proportions relating this "sequence" will have to be either 0 or 1. In other words, if the sequence of "random" determinations consists of just one element, this element will have the property of a sure, not a random, object, in the usual sense of the word.

Now let us turn to the passage from Fisher's paper quoted above, pp. 237–8, and try to see exactly what is supposed to be random there and what elements of the problem are treated as sure numbers or sure functions. These details in the set-up are not stated at the outset, but there is no difficulty in collecting them from appropriate passages in the paper. We first see that the function t of (10) is supposed to be "distributed in different samples . . . ." This means that t is a random variable and that its randomness depends on what is found in those repeated samples, namely, the values of $\bar{x}$ and s. It follows that the probabilities concerning $\bar{x}$, s, and t refer to the sequence S of those "different" samples. The sequence could not consist of just one sample because, in such a case, the "distribution" of t would not be anything like "Student's" law. The references to a normal population sampled and to "Student's" law indicate, on the contrary, that the sequence S of samples is very large indeed, and that the distributions in it are comparable to those represented by continuous curves.

Up to this time we have not mentioned the population mean μ which is also involved in the expression of t. Obviously, this may be treated mathematically either as a random or as a sure number. Both methods of approach are at our disposal but, in order to avoid inconsistencies, we must be clear as to which one we follow. The indication of Fisher's choice is found a little further on in this article, in the place describing the distinction between the fiducial and the inverse probability approach: "It is of some importance to distinguish such (fiducial) probability statements about the value of μ, from those that would be derived by the method of inverse probability from any postulated knowledge of the distribution of μ in the different populations which might have been sampled." This sentence does not seem to leave any ground for doubt. In the fiducial approach we consider but one population sampled and no distribution of μ is postulated. Therefore, μ is a sure number and, if t is distributed according to "Student's" law, it is a result of the appropriate variability of $\bar{x}$ and s alone.

The symbol t_1, which also comes into play, is obviously a sure variable capable of any real value between $-\infty$ and $+\infty$. We may select it as we wish and then obtain the probability $P(t_1)$ of the random variable t exceeding t_1 from tables.

Following the article, we will readily agree with Fisher that the inequality (11), namely, $\mu < \bar{x} - st_1/\sqrt{n}$, is equivalent to $t > t_1$ and that it must be satisfied with some probability $P(t_1)$. Now consider the phrase: "Since, therefore, the right-hand side of the inequality (i.e. $\bar{x} - st_1/\sqrt{n}$) takes, by varying t_1, all real values, we may state the probability that μ is less than any assigned value, or the probability that it lies between any assigned values, or, in short, its probability distribution *in the light of the sample observed.*" From the point of view of ordinary logic and of ordinary theory of probability this phrase is inconsistent with the original set-up. The first inconsistency is involved in the words which are italicized, suggesting that $\bar{x}$ and s in the expression $\bar{x} - st_1/\sqrt{n}$ are not random but sure numbers, referring to one particular observed sample. As a matter of fact this same inconsistency appears earlier in the statement that $\bar{x} - st_1/\sqrt{n}$, by varying t_1, will run through all real numbers. If, as formerly, $\bar{x}$ and s are random with their variation appropriate to the sequence S, then, whatever value we choose to ascribe to t_1, say $t = 2$, the expression $\bar{x} - 2s/\sqrt{n}$ is also random and depends on the outcome of sampling.

Apart from this sudden shift in the meaning ascribed to $\bar{x}$ and s, there are two more inconsistencies. To see the first of them, let us follow Fisher, changing our minds about $\bar{x}$ and s and considering them as sure numbers, determined by one particular sample. In this case the inequality $\mu < \bar{x} - st_1/\sqrt{n}$ would contain no random elements at all: the first element, μ, is an unknown constant, the mean of a single population sampled, $\bar{x}$ and s are fixed by the sample observed, and t_1 is the value of the sure variable that we have chosen to consider. In these circumstances, the inequality may either be true or not true and the probability of its being true will equal unity or zero and have nothing to do with the probability or frequency $P(t_1)$ which this same inequality satisfies within a sequence S of many "different" samples.

The last inconsistency refers, of course, to the point of view on μ. As we have seen above, it is first considered as a sure number, but the passage just quoted speaks of the probability of its lying between any assigned limits possible to determine from the values of $P(t)$. Assume $n = 4$ and that the sample observed gives $x = 10$ and $s = 2$. Select $t_1 = 0.765$ and $t_1' = -0.765$ so that $P(t_1) = 0.25$ and $P(t_1') = 0.75$. This would result in the supposed probability P' of μ lying between the limits $9.235 \leq \mu < 10.765$, being equal to $\frac{1}{2}$. Trying to interpret this result in the light of the classical theory of probability, we have to conceive a sequence, say S', of cases in 50% of which

μ falls between the above limits. But exactly what could this sequence be? Either there is such a sequence and then we must also consider other populations "which might have been sampled," *and postulate something about the distribution of* μ,[5] or else the "sequence" must be the degenerate one of one element only with the probability P' equal to either zero or unity, but never to $\frac{1}{2}$.

These are the points previously mentioned by the author (Neyman, 1934), which, from the point of view of classical probability, represent conceptual inconsistencies. They are also present in the other passage of Fisher quoted on p. 241, but a similar analysis of that passage, supplemented by what has subsequently been done by Sukhatme, will reveal errors in algebra of probability laws as well. These errors are particularly relevant from the point of view of the controversies between Bartlett and Fisher.

The quantities considered in this passage are all dependent on the population means μ and μ' and on the statistics $\bar{x}$ and s of one random sample and on $\bar{x}'$ and s' of the other. Our analysis will also require the consideration of the population variances σ^2 and σ'^2. We must start by deciding on the random or sure character of all these quantities. Fisher's remark that the two ratios

$$t = \frac{\mu - \bar{x}}{s} \quad \text{and} \quad t' = \frac{\mu' - \bar{x}'}{s'} \tag{37}$$

are distributed according to "Student's" law with appropriate degrees of freedom suggests that μ and μ' are treated as sure numbers and that $\bar{x}$, $\bar{x}'$, s, and s' are random. There is no reference whatever to the variances σ^2 and σ'^2. As nothing is disclosed about what distribution they may possess, by analogy with the μ's it is natural to treat them as sure numbers also.

In order to interpret every step in calculations more easily, we shall imagine two normal populations π_1 and π_2 sampled and a sequence A of pairs of samples, of n and n' individuals respectively, drawn independently from π_1 and π_2. These pairs of samples will determine $\bar{x}$, s, $\bar{x}'$, and s', generating distributions appropriate to normal populations. Substituted into formulae (37) they will make t and t' vary to generate the two distributions of "Student."

With this in mind, let us examine the passage in which Fisher writes

$$\epsilon = \delta - d = s't' - st, \tag{38}$$

and comments: "Since s' and s are known, the quantity represented on the right has a known distribution, though not one which has been fully tabulated." We see here the same kind of sudden jump in the point of view on quantities considered as is found in the passage analyzed previously. Formerly s' and s were not "known" but random. Otherwise, the distributions

[5] This is quite essential. Otherwise there would be an error in Bayes's theorem.

of t and t' would not have been those of "Student" but would have been normal about zero and due solely to the variability of $\bar{x}$ and $\bar{x}'$. Now s' and s are known sure numbers. Let us allow for this shift in conditions and try to visualize the character of the distribution of ϵ for fixed s' and s. For this purpose we have to consider not the whole sequence A of pairs of samples mentioned above, but only a subsequence B composed only of those pairs of samples in which the estimated variances have the same values s and s' as the ones supposed to be "known." The variability of ϵ in the subsequence B will be the result of the variability of $\bar{x}$ and $\bar{x}'$ only. It is known that the mean of a sample from a normal population is independent of the sample variance. Consequently the distributions of $\bar{x}$ and $\bar{x}'$ in B will be normal. As the connexion between ϵ on one hand and $\bar{x}$ and $\bar{x}'$ on the other is linear with constant coefficients, it would follow that the distribution of ϵ in B would be normal also. Therefore, it is with some surprise that one reads Fisher's suggestion that this distribution has not been fully tabulated. Evidently, when writing the sentence quoted, Fisher had something else in mind, probably depending on the new extra-logical principle described in subsequent publications. However this may be, we have to note the conflict between the sentence quoted and the rules of ordinary logic and of the classical theory of probability.

The distribution of ϵ by itself does not play any further role in Fisher's work. Instead he and, subsequently, Sukhatme consider the ratio that we will denote by $z = \epsilon/\sqrt{s^2 + s'^2}$. Fisher does not write any formula representing the supposed distribution of z and we have to look for the details of his ideas in Sukhatme's paper. Complimentary references to this paper in subsequent publications by Fisher suggest that it is perfectly in line with his own ideas. We quote the relevant sentence in Sukhatme's paper, only altering his notation to bring it into agreement with that of Fisher.

He (Fisher) considers the distribution of

$$z = \frac{\epsilon}{\sqrt{s^2 + s'^2}} = t' \cos R - t \sin R, \tag{39}$$

for given n, n', and R in order to obtain the probability that z exceeds any given value.

It is obvious at once that the probability in question does not refer to either of the sequences A or B visualized above. The appropriate sequence C of pairs of samples to which this probability refers is a part of the sequence A composed of all such pairs of samples in which the variances s^2 and s'^2, while variable, keep the ratio $s/s' = \tan R = \text{constant}$. Mathematically, the distribution sought is known as the relative distribution law of z given R and is denoted by $p(z \mid R)$. If $p(R)$ and $p(z, R)$ are the absolute probability law of R and the absolute joint probability law of z and R, respectively, then, for every R such that $p(R) > 0$,

$$p(z \mid R) = \frac{p(z, R)}{p(R)}. \tag{40}$$

The relative probability, given R, of z exceeding a fixed number z_1, that is, $P(z > z_1 \mid R)$, will be obtained by integrating (40) for z from z_1 to $+\infty$. There is an alternative way of obtaining the same probability. This consists of first finding the relative joint probability law given R of t and t'. If this is denoted by $p(t, t' \mid R)$ then

$$P\{z > z_1 \mid R\} = \int\int_{w(z_1)} p(t, t' \mid R)\, dt\, dt', \tag{41}$$

where the region of integration $w(z_1)$ is determined by the inequality

$$z = t' \cos R - t \sin R > z_1. \tag{42}$$

A familiar formula gives

$$p(t, t' \mid R) = \frac{p(t, t', R)}{p(R)}. \tag{43}$$

Whichever way, (40) or (43), is preferred, the resulting probability $P\{z > z_1 \mid R\}$ will have the same value and will refer to the sequence C described above.

Sukhatme has chosen to apply a quadrature procedure to calculate the integral (41) with the integrand equal to the product of two of "Student's" distributions with $n - 1$ and $n' - 1$ degrees of freedom respectively. This is just the error in algebra of probability laws mentioned above. The t and t' are distributed independently and in accordance with "Student's" laws only in the sequence A where both the means $\bar{x}$ and $\bar{x}'$ and also the variances s^2 and s'^2 are undisturbed in their random and independent variation appropriate to samples from normal populations. When calculating the probability "for a given R," we do not consider the sequence A but only its part C so selected that the ratio s/s' is constant. This selection disturbs the original distribution of s and s' and is reflected in the resulting joint distribution of t and t'.

In our calculations above (26) we have used the letters u and v for what is here denoted by t and t'. Consequently, the joint probability law $p(t, t', R)$ is obtained from (32) by merely substituting t for u and t' for v. The absolute probability law of R is easily obtained by integrating (34) with respect to z between the limits $-\infty$ and $+\infty$. The result is

$$p(R) = c\rho^{n'-1} \frac{\sin^{n-2} R \cos^{n'-2} R}{\{n(n-1)\sin^2 R + n'(n'-1)\rho^2 \cos^2 R\}^{(n+n'-2)/2}}, \tag{44}$$

with c denoting a numerical constant. Substituting (32) and (44) into (43) we obtain

$$p(t, t' \mid R) = \frac{\phi(R, \rho)}{\{n(t^2 + n - 1)\sin^2 R + n'(t'^2 + n' - 1)\rho^2 \cos^2 R\}^{(n+n')/2}}, \quad (45)$$

with $\phi(R, \rho)$ denoting a function of R, ρ, n and n' only. $p(t, t' \mid R)$ is just the function to be integrated to obtain the relative probability given R of t and t' to verify any inequality such as $t' \cos R - t \sin R > z_1$. As one would expect $p(t, t' \mid R)$ appears to depend not only on R but also on the ratio of the population variances ρ^2.

It follows that, from the point of view of the ordinary theory of probability, the Fisher-Sukhatme solution is wrong. The error consists in their confusing the absolute probability law of t and t', obtainable by integrating (32) for R, with the relative probability law given R of the same variables as given by (45). Some such error seems to have been suspected by Bartlett. Repeated denials and the reference to the extra-logical principle underlying the fiducial theory lead us to believe that from the point of view of that particular theory the error is non-existent. While accepting these explanations we may still regret that the earlier papers by Fisher and that of Sukhatme do not contain any clue as to how they are to be interpreted.

6. SUMMARY

1. The theories of fiducial argument and of confidence intervals differ in their basic conceptions. The validity of the former requires, at least in some cases, the fulfilment of various restrictions of which the theory of confidence intervals is totally free, and/or the acceptance of some new principles impossible to deduce by the rules of ordinary logic (Yates, 1939; Fisher, 1939*b*).

2. The two theories may occasionally give the same numerical results in the form of fiducial limits on one side and of confidence limits on the other. The problem of estimating the difference of means of two unknown normal populations shows, however, that this need not always be the case and that fiducial limits need not satisfy the definition of confidence limits.

3. Bartlett's criticisms of Fisher's solution of the problem just mentioned seem to be due to his considering the problem from the point of view of ordinary theory of probability and ordinary logic. In this light Fisher's solution does contain both conceptual misunderstandings (originally pointed out in the author's paper of 1934) inherent in the very concept of fiducial distribution of a parameter, and errors in algebra of probability laws. Since the first references to the new principles outside of ordinary logic, which supposedly justify the fiducial theory, were published *after* the publication of Bartlett's criticisms, the latter seem to be perfectly justified and useful.

4. Owing to a certain flaw in the ideas underlying the fiducial theory which is noticeable in passages quoted in § 4, it is impossible to insist on

any definite attitude towards it, except that of doubt. It may be useful, however, to express the following conjectures which seem to be very probable. If they are wrong then they will be put right and, as a result, the situation will be clarified.

The present author is inclined to think that the literature on the theory of fiducial argument was born out of ideas similar to those underlying the theory of confidence intervals. These ideas, however, seem to have been too vague to crystallize into a mathematical theory. Instead they resulted in misconceptions of "fiducial probability" and "fiducial distribution of a parameter" which seem to involve intrinsic inconsistencies as described in § 5. In this light, the theory of fiducial inference is simply non-existent in the same sense as, for example, a theory of numbers defined by mutually contradictory definitions.

In earlier stages when the problems treated were very simple, the fallacy involved in "fiducial probability" was not apparent. Later on, however, difficulties appeared and the new principle "which cannot be deduced by logic" seems to have been invented to disentangle them in one particular case. But the word "principle" implies some generality, hence the drift in comments on the same subjects treated in 1936 and again in 1939. From the point of view of the direction of this drift it is perhaps significant that Yates speaks of "fiducial statements" possible to make on the ground of probabilities *a posteriori* and that the paper by Jeffreys which professes the equivalence of fiducial theory with that of inverse probability appeared in the *Annals of Eugenics,* edited by R. A. Fisher.

However this may be, the only thing that the present author ventures to profess is that the theory of fiducial probability is distinct from that of confidence intervals.

REFERENCES

Bartlett, M. S. (1936). *Proc. Camb. Phil. Soc.* 32:560.

——— (1939). *Ann. Math. Statist.* 10:129.

Clopper, C. J. & Pearson, E. S. (1934). *Biometrika,* 26:404.

David, F. N. & Neyman, J. (1938). *Statist. Res. Mem.* 2:105.

Feller, W. (1938). *Statist. Res. Mem.* 2:117.

Fisher, R. A. (1925–34). *Statistical Methods for Research Workers.* London: Oliver and Boyd.

——— (1930). *Proc. Camb. Phil. Soc.* 26:528.

——— (1933). *Proc. Roy. Soc.* A, 139:343.

——— (1935). *The Design of Experiments.* London: Oliver and Boyd.

——— (1936). *Ann. Eugen., Lond.,* 6:391.

——— (1937). *Ann. Eugen., Lond.,* 7:370.

——— (1939*a*). *Ann. Eugen., Lond.,* 9:174.

——— (1939*b*). *Ann. Math. Statist.* 10:383.

Fréchet, M. (1937). *Recherches théoriques modernes sur la théorie des probabilités.* Paris: Gauthier-Villars.

JEFFREYS, H. (1939). *Theory of Probability.* Oxford: Clarendon Press.
——— (1940). *Ann. Eugen., Lond.,* 10:48.
MISES, RICHARD v. (1939). *Probability, Statistics and Truth.* London: W. Hodge and Co.
NEYMAN, J. (1934). *J. R. Statist. Soc.* 97:558.
——— (1935*a*). *Ann. Math. Statist.* 6:111.
——— (1935*b*). *Bull. Soc. Math. Fr.* 63:246.
——— (1937). *Philos. Trans.* A, 236:333.
——— (1938*a*). Lectures and Conferences on Mathematical Statistics. Graduate School, U. S. Department of Agriculture, Washington, D. C.
——— (1938*b*). *Actualités Sci. Industr.* No. 739, p. 25.
NEYMAN, J. & PEARSON, E. S. (1933). *Philos. Trans.* A, 231:289.
PEARSON, E. S. (1939). *Biometrika,* 30:471.
PEARSON, KARL (1938). *The Grammar of Science.* London: Everyman's Library.
PITMAN, E. J. G. (1939). *Biometrika,* 30:391.
PYTKOWSKI, W. (1932). *The Dependence of Income of Small Farms upon their Area, the Outlay and the Capital Invested in Cows.* Warsaw: Series Bibljoteka Pulawska, 34.
STARKEY, DAISY M. (1938). *Ann. Math. Statist.* 9:201.
"STUDENT" (1908). *Biometrika,* 6:1.
——— (1925). *Metron,* 5:18.
SUKHATME, P. V. (1938). *Sankhyā,* 4:39.
WALD, A. (1939). *Ann. Math. Statist.* 10:299.
WALD, A. & WOLFOWITZ, J. (1939). *Ann. Math. Statist.* 10:105.
WELCH, B. L. (1938). *Biometrika,* 29:350.
——— (1939). *Ann. Math. Statist.* 10:58.
YATES, F. (1939). *Proc. Camb. Phil. Soc.* 35:579.

Part 4. Stein's Sequential Procedure

(Based on a lecture at the Department of Statistics, University College, London, delivered in March, 1950.)

When Professor Egon S. Pearson invited me to speak to you, he suggested that I describe some of the more outstanding results obtained in the United States during the last decade, which, because of war conditions, may not have received the attention that they deserve. As far as I can see, the most interesting result of this description is due to Charles M. Stein. With Professor Pearson's and your permission, the subject of my today's talk will be a brief account of Stein's Sequential Procedure in estimating the mean of a normal distribution. Stein's paper [1] was published in 1945. However, in order to appreciate fully his result and, also, in order to give due credit to another friend of mine, Dr. Joseph Berkson, I shall begin my story a little earlier.

As you know, one of the earliest results in the theory of confidence intervals is the short unbiased confidence interval for the mean ξ of a

[1] Charles M. Stein: "Two-sample test of a linear hypothesis whose power is independent of the variance." *Annals of Math. Stat.,* Vol. 16 (1945), pp. 243–258.

normal distribution with an unknown variance σ^2. If $\bar{x}$ and s^2 stand for the sample mean and for the estimate of variance of this mean, based on f degrees of freedom, then the confidence interval of the unknown mean ξ is given by

$$\bar{x} - st \leqq \xi \leqq \bar{x} + st \tag{1}$$

where t is taken from Fisher's tables in accordance with the selected confidence coefficient α and the number of degrees of freedom f. I have been describing this result in my lectures since about 1930, and it was first used by W. Pytkowski [2] in his booklet published in 1932. The theoretical background is given in my *J.R.S.S.* paper of 1934. Finally, the corresponding result based on fiducial argument was published by R. A. Fisher in 1935. Furthermore, in my paper of 1937 published in the *Phil. Trans. Roy. Soc.*, London, I have shown that the confidence interval (1) has the remarkable properties of "unbiasedness" and "shortness." "Unbiasedness" means the property that, while the true value of ξ is covered by (1) with the preassigned frequency α, any other value is covered by (1) less frequently. "Shortness" means that, given a false value ξ' of ξ, the confidence interval (1) covers ξ' less frequently than (or at most as frequently as) any other unbiased confidence interval corresponding to the same confidence coefficient α.

Analytically, these properties are expressed as follows. The property serving as the definition of a confidence interval $\{\xi_1(E), \xi_2(E)\}$ is

$$P\{\xi_1(E) \leqq \xi \leqq \xi_2(E) \mid \xi, \sigma\} \equiv \alpha. \tag{2}$$

Here, as usual, the letter E stands for the random "event" point, i.e. for the set of all the observable random variables. The property of unbiasedness is expressed by the relation

$$P\{\xi_1(E) \leqq \xi' \leqq \xi_2(E) \mid \xi, \sigma\} \leqq P\{\xi_1(E) \leqq \xi \leqq \xi_2(E) \mid \xi, \sigma\} \tag{3}$$

valid for all values of ξ, ξ' and σ. Finally, the property of shortness, applicable to (1), is written as

$$P\{\bar{x} - st \leqq \xi' \leqq \bar{x} + st \mid \xi, \sigma\} \leqq P\{\xi_1(E) \leqq \xi' \leqq \xi_2(E) \mid \xi, \sigma\} \tag{4}$$

for all confidence intervals $\{\xi_1(E), \xi_2(E)\}$ satisfying (2) and (3), and for all ξ, ξ' and σ.

I must admit that, having obtained this result, I thought that I had found a grand thing, not only interesting theoretically, but also important practically, and felt naively proud. Unfortunately, this inordinate pride was soon punctured by a letter from Dr. Berkson, expressed in polite terms but making it quite clear that the practical importance of the confidence

[2] See references in part 3 of this Chapter.

interval (1) is rather limited. The humiliating part of the story is that Joseph Berkson is an M.D. and a practical, rather than a theoretical statistician, while I am supposed to be working in theory. Yet a delicate point regarding the confidence interval (1) was noticed by Berkson and overlooked by me. In this connection it seems appropriate to paraphrase the celebrated description of Chevalier de Méré due to Pascal which, at tea time, you see on the wall of the Common Room: [3] Il n'est pas géomètre, mais il a très bon esprit et ça, comme vous savez, est un grand avantage. . . .

The practical defect of the confidence interval (1) noticed by Berkson is that its length, viz. $2st$, is a random variable and, what is more, a variable capable of assuming arbitrarily large values. In fact, by looking up Elderton's tables relating to the distribution of χ^2, it is easy to compute the probability that the length of confidence interval (1) will exceed any preassigned limit. In order to appreciate the practical importance of this fact, imagine an M.D., engaged in some sort of routine analysis, applying interval (1) to estimate, say, the average sugar content in a patient's blood. This estimate is needed in order to adjust appropriately the dose of an injection. If we grant all the approximations involved, it is obvious that frequently the particular determinations used by the M.D. will be concordant and the value of s will be small. In these cases, assertions (1) regarding the true value of ξ will be usable. However, in other cases the value of s will be large and then the assertion regarding ξ will be so vague, say from zero to 100 per cent, as to be meaningless.

Cases of this kind are, of course, familiar and you must have come across a substantial literature dealing with so-called "gross errors." Gross errors must occur from time to time. However, the situation I have in mind is not concerned with gross errors but only with such variation of the estimate s as is implied by the postulated normal distribution of the particular determinations.

Faced with the abnormal length of the confidence interval for the mean sugar content ξ, the M.D. can do only one thing: not use this confidence interval. This may be followed by taking another sample of blood and making a new series of determinations, or by computing a new confidence interval based on some of the original determinations after rejecting suspected "gross errors." But these further steps concern us less than the predominant fact that a universal application of confidence interval (1) is

[3] Since the time of Karl Pearson, the decorations of the Common Room (where a friendly visitor may get tea at 3.45 P.M., irrespective of whether he—or she—is mathematically minded or not) include the following quotation from Pascal, written in beautiful Gothic:

> "Il a très bon esprit; mais il n'est pas géomètre; c'est, comme vous savez, un grand défaut."

impractical. Its use is limited to cases where the value of the estimated standard error s does not exceed a certain (no doubt, only vaguely determined) limit τ. This, however, implies that the confidence interval (1) is not used at all.

This is a point of some delicacy and it is worthwhile to emphasize it a little. As I have already mentioned, the term "confidence interval corresponding to the confidence coefficient α" is used to describe the interval between two functions of the observable random variables $\xi_1(E)$ and $\xi_2(E)$ having the property of bracketing the true value of the estimated parameter ξ with the preassigned probability α. If we equate

$$\xi_1(E) = \bar{x} - st,$$

$$\xi_2(E) = \bar{x} + st$$

and use these two functions consistently to estimate ξ, *irrespective of the observed values* of $\bar{x}$ and s, then the long run relative frequency of successful estimates will actually be α. However, if we restrict the use of these formulae to cases where $s \leqq \tau$, then, strictly speaking, our estimating interval will not be bounded by functions $\xi_1(E)$ and $\xi_2(E)$ defined above but by two new functions, say $\xi_1^*(E)$ and $\xi_2^*(E)$ defined as follows.

$$\xi_i^*(E) \equiv \xi_i(E) \qquad \text{whenever } s \leqq \tau,$$

$$\xi_i^*(E) \qquad \text{not defined otherwise,}$$

for $i = 1, 2$. For convenience of reference, the interval (ξ_1^*, ξ_2^*) will be described as the curtailed confidence interval for ξ.

Unexpected as it may seem, the two functions $\xi_1^*(E)$ and $\xi_2^*(E)$ do not possess the properties of confidence limits, because the probability that they will bracket the true value of ξ is less than α and depends on the value of σ. Let us compute this probability, say P. This is the conditional probability, given $s \leqq \tau$, that

$$\bar{x} - st \leqq \xi \leqq \bar{x} + st$$

where ξ represents the true value of the expectation of $\bar{x}$. We have

$$P = \frac{P\{(s \leqq \tau)(\bar{x} - st \leqq \xi \leqq \bar{x} + st) \mid \xi,\sigma\}}{P\{s \leqq \tau \mid \sigma\}}. \tag{5}$$

In order to evaluate the denominator, we need the probability density function of s, say,

$$\frac{c}{\sigma^f} s^{f-1} e^{-nfs^2/2\sigma^2} \tag{6}$$

where c is a numerical factor, independent of σ.

In order to compute the numerator, in addition to (6) we need the probability density function of $\bar{x}$,

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(\bar{x}-\xi)^2/2\sigma^2}. \tag{7}$$

Owing to the independence of $\bar{x}$ and s, the joint probability density function of $\bar{x}$ and s is simply the product of (6) and (7). The numerator in (5) is

$$P\{(s \leqq \tau)(|\bar{x} - \xi| \leqq st) \mid \xi,\sigma\}$$
$$= \frac{c}{\sigma^f}\int_0^\tau \left\{ s^{f-1}e^{-nfs^2/2\sigma^2} \frac{2\sqrt{n}}{\sigma\sqrt{2\pi}} \int_\xi^{\xi+st} e^{-n(\bar{x}-\xi)^2/2\sigma^2} d\bar{x} \right\} ds.$$

Similarly,

$$P\{s \leqq \tau \mid \sigma\} = \frac{c}{\sigma^f}\int_0^\tau s^{f-1}e^{-nfs^2/2\sigma^2} ds.$$

The integrals simplify if we substitute

$$\frac{\sqrt{n}(\bar{x} - \xi)}{\sigma} = u,$$

then let

$$G(x) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2} dt \tag{8}$$

and, finally, put

$$\sqrt{n}\frac{s}{\sigma} = v.$$

Then

$$P\{(s \leqq \tau)(|\bar{x} - \xi| \leqq st) \mid \xi,\sigma\} = \frac{c}{n^{f/2}}\int_0^{\tau\sqrt{n}/\sigma} v^{f-1}e^{-fv^2/2}G(vt)\, dv,$$

$$P\{s \leqq \tau \mid \sigma\} = \frac{c}{n^{f/2}}\int_0^{\tau\sqrt{n}/\sigma} v^{f-1}e^{-fv^2/2}\, dv,$$

and

$$P = \frac{\int_0^{\tau\sqrt{n}/\sigma} v^{f-1}e^{-fv^2/2}G(vt)\, dv}{\int_0^{\tau\sqrt{n}/\sigma} v^{f-1}e^{-fv^2/2}\, dv}.$$

It is seen that P is a weighted average of quantities

$$G(vt)$$

where, for each v, the weight is represented by

$$v^{f-1}e^{-fv^2/2}.$$

Since $G(x)$ is a monotone function of x, increasing from zero to unity as x grows from zero to infinity, it is obvious that P depends on σ and, namely, that as σ is increased from zero to infinity, the value of P decreases from α to zero. Thus, if τ is fixed in one way or another, and we make assertions about ξ in the form

$$\bar{x} - st \leqq \xi \leqq \bar{x} + st \tag{9}$$

only if $s \leqq \tau$, then the probability of this assertion being correct is always less than the chosen confidence coefficient α and, if σ happens to be large, is close to zero. It follows that interval (9) used only when $s \leqq \tau$, is not a confidence interval.

As you know, the properties of confidence interval (1) are connected with Student's distribution. This has an extensive use in testing Student's hypothesis which ascribes a specific value ξ_0 to the mean ξ of the normal distribution but fails to specify the value of the standard error σ. Student's test consists of the rule to reject the hypothesis tested when the criterion

$$\frac{|\bar{x} - \xi_0|}{s}$$

exceeds a specified value t.

This test was proved [4] to be unbiased of type B1, which means that it is the most powerful test of all tests which are unbiased. Yet, you must be aware of the fact that it has an unpleasant property. This property is that the power function of this test depends on the unknown value of σ. In fact, the argument of the power function is

$$\rho = \frac{|\xi_0 - \xi|}{\sigma}\sqrt{n},$$

where ξ stands for the true value of the mean. As ρ is increased, the power function tends to unity and there are some tables from which its values can be read. One of the uses for which these tables are intended is to estimate how large should n be in order to have a reasonable chance of detecting the falsehood of the hypothesis when the true mean ξ differs from the hypothetical value ξ_0 by a stated amount. Upon inspecting the expression for the argument of the power function you will see that, when nothing is known about σ, it is impossible to answer this question. In fact, however large be n, if σ is sufficiently large, then ρ will be as small as desired and the value of the power function close to the chosen level of significance.

[4] J. Neyman: "Sur le vérification des hypothèses statistiques composées." *Bull. Soc. Math. de France,* t. 63 (1935), pp. 246–266.

When we have some knowledge of σ, for example if we know that σ cannot exceed a specified limit, then the tables of the power function of Student's test can be used to estimate the upper bound of n needed to insure that the power function does not fall below a desired level. Similarly, if we know the upper bound of σ, we can select sufficiently large values of $\tau\sqrt{n}$ so that the probability P of success in estimating ξ using the curtailed confidence interval be at least equal to a specified value α. In many cases the value of σ is not entirely unknown and then both the power function and the curtailed confidence interval for ξ are usable. Otherwise, we face a very unpleasant difficulty.

As you know, the properties of a test are determined by the corresponding critical region. Similarly, the properties of a confidence interval are those of the corresponding regions of acceptance.

With this in mind, a question occurred to me as to whether or not it is possible to find critical regions for testing Student's hypothesis and regions of acceptance for estimating ξ having more satisfactory properties. From the critical region, w, we would require that it correspond to a preassigned level of significance,

$$P\{E \epsilon w \mid \xi_0, \sigma\} \equiv \epsilon$$

and that for sufficiently large values of $\mid \xi_0 - \xi \mid$, the power function $\beta(\xi, \sigma \mid w)$ of the region w have sufficiently large values irrespective of the value of σ. Of course, it would be most satisfactory if $\beta(\xi, \sigma \mid w)$ were independent of σ and could tend to unity as $\mid \xi_0 - \xi \mid$ is increased. As regards the regions of acceptance, we would require that they correspond to the preassigned confidence coefficient α and that the length of corresponding confidence intervals never exceed a fixed finite number M.

It is easy to see that there is a connection between the two questions. In fact, a negative answer to the question regarding the critical region implies a negative answer to the question regarding the regions of acceptance. To see this, assume for a moment that a system A of regions of acceptance $A(\xi)$ is found, corresponding to the confidence coefficient $\alpha = 1 - \epsilon$, such that the length of the corresponding confidence intervals does not exceed M. We shall see that this assumption implies the existence of a critical region w corresponding to the level of significance ϵ and such that, whenever $\mid \xi_0 - \xi \mid > M$, the power function

$$\beta(\xi, \sigma \mid w) \geqq 1 - \epsilon$$

irrespective of the value of σ.

In order to prove this proposition, notice that, if $\mid \xi_0 - \xi \mid > M$, then no confidence interval can cover both ξ and ξ_0. This means that the region of acceptance $A(\xi)$ and the region of acceptance $A(\xi_0)$ have no points in common. In other words, $A(\xi)$ lies entirely within the region $w = W - A(\xi_0)$.

Now select the region $w = W - A(\xi_0)$ as the critical region for testing Student's hypothesis that ascribes to ξ the value ξ_0. Using the basic property of the region of acceptance, we have

$$P\{E \epsilon A(\xi_0) \mid \xi_0, \sigma\} \equiv \alpha = 1 - \epsilon.$$

Hence

$$P\{E \epsilon w \mid \xi_0, \sigma\} \equiv 1 - \alpha = \epsilon.$$

Thus, the region w corresponds to the preassigned level of significance. Assume now that the hypothesis tested is false and that the true value ξ differs from ξ_0 by more than M. The value of the power function corresponding to this value is

$$\beta\{\xi, \sigma \mid w) = P\{E \epsilon w \mid \xi, \sigma\}.$$

But, as we have noticed above, the region w includes $A(\xi)$. Hence

$$\beta(\xi, \sigma \mid w) \geqq P\{E \epsilon A(\xi) \mid \xi, \sigma\} \equiv 1 - \epsilon,$$

irrespective of the value of σ. Q.E.D.

The questions just described were attacked by Dr. George B. Dantzig, then a colleague of mine, and were answered in the negative. Studying the structure of regions similar to the sample space with regard to σ, while $\xi = \xi_0$ is kept fixed, Dantzig found that, if the power function of such a region is independent of σ, then the region is similar to W not only with respect to σ, but also with respect to ξ and, therefore,

$$\beta(\xi, \sigma \mid w) \equiv \epsilon$$

identically in ξ and σ. This result appeared in print.[5] Furthermore, Dantzig proved a more general proposition: *Whatever be the region w, similar to the sample space with respect to σ, when $\xi = \xi_0$, and whatever be $\xi_1 \neq \xi_0$, the upper limit of its power function $\beta(\xi, \sigma \mid w)$ as $\sigma \to \infty$ cannot exceed ϵ.* The proof of this proposition is very simple. It is known [6] that the asymmetric Student's test has the property of being the uniformly most powerful test of Student's hypothesis tested against the set of admissible hypotheses ascribing to the mean ξ values on one side of the hypothetical value ξ_0. Assume, for example, that $\xi_1 > \xi_0$. Then the most powerful test of the hypothesis that $\xi = \xi_0$ tested against the alternative $\xi = \xi_1$ has its critical region, say w, defined by the inequality,

$$\bar{x} > \xi_0 + st(\epsilon),$$

[5] George B. Dantzig: "On the non-existence of tests of Student's hypothesis having power functions independent of σ." *Annals of Math. Stat.*, Vol. 11 (1940), pp. 186–192.

[6] J. Neyman and E. S. Pearson: "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc., London,* Ser. A, Vol. 231 (1933), pp. 289–337.

where $t(\epsilon)$ is a suitable constant. It follows that

$$\beta(\xi_1, \sigma \mid w_0) \geqq \beta(\xi_1, \sigma \mid w).$$

However, it is known that

$$\lim_{\sigma \to \infty} \beta(\xi_1, \sigma \mid w_0) = \epsilon,$$

and it follows that

$$\sup \lim_{\sigma \to \infty} \beta(\xi_1, \sigma \mid w) \leqq \epsilon.$$

Thus, within the n-dimensioned space W there are no "satisfactory" critical regions for testing Student's hypothesis and, consequently, there are no systems of regions of acceptance which generate confidence intervals whose length is bounded, i.e. does not exceed a fixed number M.

As you see, the situation is unsatisfactory. It was in this unsatisfactory state that it was faced by Stein, then in the United States Army. He had been assigned to study some statistical problems connected with weather forecasting and, hence, forced to learn some theory of statistics.

There were, among the things that Stein read, the now celebrated papers of Abraham Wald, dealing with so-called "sequential analysis," which, however, seems to be more appropriately called "sequential procedures." Generally, a sequential procedure in testing a statistical hypothesis consists of a repeated application of a triple rule: (a) to reject the hypothesis tested on data available, (b) to accept it on the same data or (c) to make a specified number of fresh observations. You begin by observing, say, n_1 random variables X_1, $X_2, \cdots, X_{n_1}$. The totality of these observations is represented by the sample point E_1, in the n_1 dimensioned space W_1. The space W_1 is divided into three parts, $W_1(a)$, $W_1(b)$, $W_1(c)$, and, following the determination of E_1, the statistician takes action a, b or c according to whether E_1 falls in $W_1(a)$, in $W_1(b)$ or in $W_1(c)$. In the latter case, he makes n_2 fresh observations X_{n_1+1}, X_{n_1+2}, $\cdots$, $X_{n_1+n_2}$. This number n_2 may be preassigned or, again, it may be a random variable, a function of E_1. If n_2 is a fixed constant, then the n_2 new observations combine with the original n_1 to determine a point, say E_2, in the $n_1 + n_2$ dimensioned space W_2. If n_2 is a random variable capable of assuming arbitrarily large values, then, in order to "accommodate" the sample point E_2 it is necessary to consider the space of infinitely many dimensions. This is also true in the frequent case where at every stage of sampling there exist possible sample points at which the statistician will take more and more observations.

Early writings of Wald and of his colleagues were mostly concerned with sequential sampling procedures of testing a simple hypothesis against a single simple alternative. However, these writings inspired Stein with the idea that spaces with infinitely many dimensions are somewhat "wider" than spaces

of a finite number of dimensions. Hence, if in the spaces of finitely many dimensions there are no desirable critical regions for testing Student's hypothesis nor desirable regions of acceptance for estimating the mean of a normal distribution, such regions may exist in the space of infinitely many dimensions. After some effort, Stein invented a two-step sequential procedure which proved that his presumption was correct.

Using this procedure we obtain a test of Student's hypothesis corresponding to a preassigned level of significance ϵ, with a power function independent from σ and tending to unity as the "error" of the hypothesis tested is increased. Moreover, for any given $\xi \neq \xi_0$ it is possible to arrange that the power function at the point ξ be equal to a preassigned value $\beta > 0$, as close to unity as desired.

The same sequential procedure leads to confidence intervals for the estimated ξ which both correspond to a preassigned confidence coefficient α and have a preassigned length 2Δ. The originality of the idea and the elegance of the solution are above all praise.

I shall begin by explaining Stein's procedure of obtaining the confidence interval. Next I shall show that it has the properties indicated. Thereafter the procedure of testing Student's hypothesis will be more or less evident.

Let α and 2Δ denote, respectively, the preassigned confidence coefficient and the preassigned length of the confidence interval. Stein's procedure consists of making two sets of observations. The first set of an arbitrary number $n_1 \geqq 2$ of observations

$$X_1, X_2, \cdots, X_{n_1}$$

is obtained and certain calculations are made. The result of these calculations determines the number $n_2 \geqq 1$ of observations

$$X_{n_1+1}, X_{n_1+2}, \cdots, X_{n_1+n_2}$$

of the second set. Then the two sets are combined to determine the confidence interval for the unknown mean ξ of the normal distribution sampled.

The calculations relating to the first set of observations, leading to the value of n_2 are as follows. Denote by $\tau(\alpha)$ the value of Fisher's t corresponding to the number of degrees of freedom $n_1 - 1$ and to the confidence coefficient α. In other words, $\tau(\alpha)$ is the root of the equation

$$\int_0^{\tau(\alpha)} \frac{dt}{\left(1 + \frac{t^2}{n_1 - 1}\right)^{n_1/2}} = \alpha \int_0^{\infty} \frac{dt}{\left(1 + \frac{t^2}{n_1 - 1}\right)^{n_1/2}}.$$

Having read $\tau(\alpha)$ from Fisher's tables, we compute the expression

$$\theta = \left(\frac{\tau(\alpha) S}{\Delta}\right)^2 \tag{10}$$

where S^2 is the estimate of variance computed from the first set of observations

$$S^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2$$

with

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i.$$

Now we determine n_2. If θ is less than n_1, then we put $n_2 = 1$. Otherwise, if $\theta \geqq n_1$, then n_2 is given the value equal to the least integer which exceeds $\theta - n_1$. It will be seen that in either case

$$n_1 + n_2 > \theta. \tag{11}$$

We shall need this inequality at a later stage. It will be seen that the greater S, the greater the value of n_2.

When the value of n_2 is determined, we make the second set of observations and compute the corresponding mean, say

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} X_i.$$

Stein's confidence interval for ξ is then given by the following formula

$$a\bar{X}_1 + (1 - a)\bar{X}_2 - \Delta \leqq \xi \leqq a\bar{X}_1 + (1 - a)\bar{X}_2 + \Delta \tag{12}$$

where the value of a is obtained from the equation

$$\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2} = \frac{1}{\theta}. \tag{13}$$

You will observe that the difference between the extreme parts of (12) is equal to 2Δ. Thus, the only thing which requires proof is that, given that the mean of the sampled normal distribution is ξ, the probability

$$P\{a\bar{X}_1 + (1 - a)\bar{X}_2 - \Delta \leqq \xi \leqq a\bar{X}_1 + (1 - a)\bar{X}_2 + \Delta \mid \xi, \sigma\} \equiv \alpha.$$

This identity can be rewritten as

$$P\{\left| a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi \right| \leqq \Delta \mid \xi, \sigma\} \equiv \alpha. \tag{14}$$

In order to prove (14) we first verify that, with the described selection of the value of n_2, equation (13) has real roots. Upon multiplying by $n_1 n_2 \theta$ and sorting out terms, this equation may be rewritten as

$$(n_1 + n_2)\theta a^2 - 2n_1\theta a + n_1(\theta - n_2) = 0$$

and the two roots are

$$a_{1,2} = \frac{n_1\theta \pm \sqrt{n_1 n_2 \theta(n_1 + n_2 - \theta)}}{(n_1 + n_2)\theta}.$$

Because of (11) the roots are real. Either one can be used in (12).

The proof of (14) is based on the following elements. The mean $\overline{X}_1$ is independent of S and is normally distributed about ξ with variance σ^2/n_1. The number n_2 is a single valued function of S, and hence a random variable. So also is a. The mean $\overline{X}_2$ depends on S only through the number n_2 of observations on which it is based. Thus, for a given S, the conditional distribution of, say,

$$X = a\overline{X}_1 + (1 - a)\overline{X}_2$$

is normal, with expectation

$$E(X) = a\xi + (1 - a)\xi \equiv \xi$$

and variance

$$\sigma^2_{x \mid s} = \left(\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}\right)\sigma^2.$$

It follows that the conditional distribution, given S, of

$$\frac{a\overline{X}_1 + (1 - a)\overline{X}_2 - \xi}{\sqrt{\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}}}$$

is normal about zero with variance σ^2. A further consequence is that the absolute distribution of the quotient

$$\frac{a\overline{X}_1 + (1 - a)\overline{X}_2 - \xi}{S\sqrt{\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}}}$$

is Student's distribution with $n_1 - 1$ degrees of freedom. Thus recalling the definition of $\tau(\alpha)$, we have

$$P\left\{\frac{\left| a\overline{X}_1 + (1 - a)\overline{X}_2 - \xi \right|}{S\sqrt{\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}}} \leqq \tau(\alpha) \mid \xi, \sigma\right\} \equiv \alpha. \tag{15}$$

However, because of (13) and (10) we have

$$S\sqrt{\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}} = \frac{\Delta}{\tau(\alpha)},$$

and it is seen that (15) coincides with (14). This completes the proof of the assertion that (12) represents a confidence interval corresponding to the confidence coefficient α.

As to Stein's test of Student's hypothesis, formula (14) is very suggestive. As you must have guessed, if ξ_0 is the value of the mean specified by the hypothesis, the test criterion is, say

$$Y = a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi_0,$$

and the symmetric test consists in the rule of rejecting the hypothesis whenever $| Y |$ exceeds Δ. According to formula (14), this test corresponds to the level of significance $\epsilon = 1 - \alpha$.

The test just described contains two arbitrary elements. One of them is the level of significance ϵ and its choice must be governed by considerations of the importance of avoiding errors of the first kind. The choice of ϵ determines α and hence $\tau(\alpha)$. The second arbitrary element in the procedure is Δ. In the theory of Stein's confidence interval, Δ plays an independent role. It represents one-half of the length of the confidence interval and is selected as such. I shall now show that in the theory of Stein's test of Student's hypothesis, the arbitrariness of Δ may be used to insure that for a given value of the difference $| \xi_0 - \xi |$ the power function of the test has a preassigned value β.

Denote by $\beta(\xi \mid \Delta)$ the power function of Stein's test. If ξ stands for the true value of the mean, we have

$$\beta(\xi \mid \Delta) = 1 - P\{| Y | \leqq \Delta \mid \xi, \sigma\}.$$

The value of the probability in the right hand side is easily computed by noticing that

$$Y = (a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi) + (\xi - \xi_0)$$

and by recalling that

$$\frac{a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi}{S\sqrt{\frac{a^2}{n_1} + \frac{(1 - a)^2}{n_2}}} = \frac{\tau(\alpha)}{\Delta}(a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi)$$

follows the Student's distribution with $n_1 - 1$ degrees of freedom. Easy algebra gives

$$P\{| Y | \leqq \Delta \mid \xi, \sigma\} =$$

$$P\{b(\xi, \Delta) - \tau(\alpha) \leqq \frac{\tau(\alpha)}{\Delta}(a\bar{X}_1 + (1 - a)\bar{X}_2 - \xi) \leqq b(\xi, \Delta) + \tau(\alpha) \mid \xi, \sigma\}, \tag{16}$$

where, for the sake of brevity,

$$b(\xi, \Delta) = \frac{\tau(\alpha)}{\Delta}(\xi_0 - \xi).$$

It follows that $P\{|\, Y \,| \leqq \Delta \,|\, \xi, \sigma\}$ is equal to the integral of Student's probability density function with $n_1 - 1$ degrees of freedom taken over the interval of length $2\tau(\alpha)$ centered at $b(\xi, \Delta)$. Therefore the power function is

$$\beta(\xi \,|\, \Delta) = 1 - \frac{\displaystyle\int_{b-\tau}^{b+\tau} \frac{dt}{\left(1 + \frac{t^2}{n_1 - 1}\right)^{n_1/2}}}{\displaystyle\int_{-\infty}^{+\infty} \frac{dt}{\left(1 + \frac{t^2}{n_1 - 1}\right)^{n_1/2}}}.$$

Obviously, for fixed $\tau(\alpha)$ and $\xi \neq \xi_0$, the value of $|\, b(\xi, \Delta) \,|$ is close to zero when Δ is large and goes to infinity as Δ decreases. At the same time the value of $\beta(\xi \,|\, \Delta)$ varies continuously from ϵ to unity. Thus, for given values ξ and β, the value of Δ can be adjusted so that $\beta(\xi \,|\, \Delta) = \beta$. Q.E.D.

Casual inspection of formula (16) may cause a sensation of surprise at the conclusion just reached. However, this sensation disappears when one recalls that a change in the value of Δ makes a change in the value of

$$\theta = \left(\frac{\tau(\alpha)S}{\Delta}\right)^2$$

and this, in turn, influences the number n_2 of observations of the second set on which the mean $\overline{X}_2$ is calculated. The greater the desired power corresponding to a given ξ, the smaller must be Δ, the larger θ corresponding to an observed S, and the larger n_2. Thus, with Stein's procedure, we can preassign both the level of significance $\epsilon = 1 - \alpha$ and the power corresponding to a chosen size of error in the hypothesis, $|\, \xi_0 - \xi \,|$. However, the more exigent we are in either respect, the more observations will be needed to achieve the desired goal.

The above account of Stein's work does not cover all of his results and, if you study his paper, you will find it interesting and informative. Among other things you will find in it the description of another procedure, slightly more efficient than that described, and a generalization of these results to the case of the general linear hypothesis.

The most essential advance achieved, as I presented it, consists in the shift from studies of sample spaces having finitely many dimensions to studies of the sample space of infinitely many dimensions, and the proof

that in the latter there are various possibilities which are not available in the former. Stein proved this point by giving an ingenious example. Now the door is open for a search for an optimum sequential procedure. This problem appears rather difficult, but Stein has already obtained some relevant results which will soon appear in print.

I began this lecture with a reference to some correspondence with Joseph Berkson in which he complained that, since the original confidence interval (1) for estimating the mean of a normal distribution is unbounded and, from time to time, must be inordinately long, a consistent use of this interval in practical work is impossible.

Now, this difficulty seems to have been removed by means of Stein's work. Brilliant as his result is, we must realize that its practical applications involve a new difficulty, just as insuperable as that complained of by Berkson. This difficulty is connected with the fact that in the course of repeated attempts to apply Stein's procedure, the observed value of S will be exceedingly large from time to time and will determine a correspondingly large value of n_2. Likewise, it is obvious that, if the M.D., of whom I spoke at the beginning of this lecture, is advised to make an additional $n_2 = 1{,}000{,}000$ determinations of sugar in a patient's blood, he will refuse. Thus, in all practical work it will be unavoidable to apply some sort of "curtailed" Stein procedure. However, this conclusion need not inspire us with undue pessimism. A strict accordance between practical work and a corresponding theory is never possible and yet all our life is based on constant practical applications of inapplicable theories. For example, we postulate that the M.D.'s analyses follow a normal law of frequency whereas it is quite plain that none of his determinations can be negative and none can exceed 100. By assuming normality we substitute "improbability" instead of "impossibility" and are content. So does Berkson. If he now complains of the possibility of tremendous values of n_2, we may point out that such values are extremely improbable and may advise him to be satisfied by making a reverse substitution of "impossibility" instead of "improbability."

. . . . .

Now, I wish to add a little postscript to the above lecture on Stein's results and to the whole collection of lectures and conferences assembled in this book. This postscript has to deal with the general character of statistical research and with the ties that exist between the pure mathematical theory of statistics and the applied work. I deeply regret the not infrequent emphatic declarations for or against pure theory and for or

against work in applications.[7] It is my strong belief that both are important and, certainly, both are interesting. The Berkson-Dantzig-Stein incident just recounted provides an excellent illustration of the view that, thus far, mathematical statistics is still in its early phase of development and that the various fields of applied statistical work constitute the source of interesting problems of theory. The results of Dantzig and Stein are certainly contributions to pure theory of statistics. Yet, whether the two authors are aware of the fact or not, the theoretical problems they solved originated from difficulties in applied work. Further development of mathematical statistics, and also the success of university instruction of statistics, depend upon maintaining close contact and a harmonious balance between mathematical direction of thought and the various fields of application.

[7] Quite recently I was shown some letters regarding myself. One very nice person wrote "I met Neyman. In general he is O.K., but hopelessly mathematical. . . ." The letter of another equally nice person stated: "Once upon a time Neyman did some real work. Now, however, he is interested in applications."

INDEX OF NAMES

INDEX OF TERMS