LECTURES AND CONFERENCES ON MATHEMATICAL STATISTICS

Delivered by

J. Neyman

at the Graduate School of the
United States Department of Agriculture
in April 1937

Revised and supplemented by the author
with the editorial assistance of
W. Edwards Deming

Published by

The Graduate School
of the
United States Department of Agriculture
Washington

2

Price $1.25

LECTURES AND CONFERENCES ON MATHEMATICAL STATISTICS

Delivered by

J. Neyman

at the Graduate School of the
United States Department of Agriculture
in April 1937

Revised and supplemented by the author
with the editorial assistance of
W. Edwards Deming

Published by

The Graduate School
of the
United States Department of Agriculture
Washington

Price $1.25

Mimeographed in the United States of America

PREFACE

The present record of my lectures and conferences would not be complete without the acknowledgment of my deep gratitude to the organizing committee appointed by Dr. A. F. Woods, the Director of the Graduate School, for the kind invitation extended to me to lecture in that important centre, and to the audience for the friendly reception and interest offered to my talks.

I owe a special and very warm indebtedness to Dr. W. Edwards Deming, who was kind enough to advise me on the topics that would be of interest to the audience, and who planned and thought out all the details of my one week's stay in Washington, which I so thoroughly enjoyed.

It is a great honour for me to have spoken at the Graduate School of the United States Department of Agriculture, the more so as the audience included many eminent statisticians, whose work in various directions I have greatly admired. When communicating to them the results of my own studies and those of the persons with whom I was, or still am, associated, I hoped for occasions to learn myself. These were amply provided by the discussions which followed the lectures and conferences. The questions put to me both before and after my talks, and also the critical remarks, were most interesting and suggestive. Frequently they referred to some practical or theoretical difficulties encountered in the statistical work carried out on such an imposing scale in the United States. Some of these problems were quite new to me and I was not able to offer any reply to many a question asked. Later on, however, I managed to produce some of the answers required and these, I hope, will soon be published elsewhere. Other questions and remarks suggested that some points in my lectures were not sufficiently clearly presented. Consequently, I tried to introduce the necessary amendments to make things clearer, and the present draft differs in places from what I actually said. In this particular respect I owe very much to Dr. Deming whose inquisitiveness and friendly criticism helped much in improving clearness and accuracy of presentation.

It is a pleasure to record a similar indebtedness to Mr. Milton Friedman, Dr. Charles F. Sarle, Mr. Frederick F. Stephan, Dr. Sidney Wilcox, and others, who were kind enough to lend me their attention and help.

It is again to Dr. Deming that I am most grateful for his idea of publishing the present book and for having taken infinite trouble in producing it in the excellent form in which it actually appears.

I must add one remark concerning the contents of the lectures and conferences as they appear in the present publication. It will be seen that some of them are concerned with pure theory, others with various applications: problems of plant breeding, those of randomized and systematic arrangements of agricultural trials, of sampling human populations, and of time series analysis. Needless to say, the audiences varied considerably from one lecture or conference to another. Consequently, when speaking on applications requiring a reference to certain details of the theory, I did not hesitate to mention them at some length, even if I had already had the occasion of discussing the same point at a previous lecture or conference. This was necessary because many of the listeners of one conference did not attend the others. In drafting the present publication, we have thought it best to avoid some of the repetitions that actually occurred. However, it was thought wise not to omit them altogether, since, if that had been done, each of the conferences dealing with applications to practical problems would not be a closed unit in itself.

J. Neyman.

The Cell, Little Hampden,
Great Missenden,
Buckinghamshire.
28 August 1937

FOREWORD FROM THE EDITOR

It may not be out of place to recall that the relation between editor and author is different from that between co-authors. An editor is responsible for clarity, cross-references, citations to literature, proof-reading, and general appearance, but save for notes actually signed "editor," he is not responsible for the actual content of the material, however extensively he may have revised it and contributed to it. On the other hand, co-authors are jointly accountable for every portion of an article to which their names are attached.

Anticipating that the statistical methods developed in this book will have considerable theoretical and economic value in agriculture, Dr. A. F. Woods, Director of the Graduate School, and Dr. C. H. Kunsman, Chief of the Fertilizer Research Division of the Bureau of Chemistry and Soils, have put the facilities of their offices at the disposal of the editor, including whatever portion of his time could be spared from other pursuits.

The editing of this book for Dr. Neyman has been a pleasant task, made so by the enthusiastic assistance of many friends; it could hardly have been produced except for a fortunate chain of additions of their efforts. Dr. Neyman himself devoted many days of the summer of 1937 to revising the edited record of the lectures and conferences originally delivered in Washington, and has then and since been exceedingly patient in dealing with suggestions from the editor. It is also a pleasure to record in particular the valued assistance of Messrs. B. R. Stauber, Alexander Sturges, Milton Friedman, W. Allen Wallis, Otis A. Pope, and Frederick F. Stephan, in various matters. The original recording of the conferences was done by Miss Helen Evans: if more expert reporters there be, the editor would shun the assignment of finding one. The typing is mostly the work of Mr. Stanley J. Magdurakas, who unacquainted with mathematics, was picking up editor's slips before the job was finished.

The primary reason for inserting this foreword was to provide space for apologies for flaws. For example, in order to make use of certain previously published graphs, some inconsistencies and infelicities in nomenclature were allowed to stand. Unfortunately the typewriter was not properly equipped with small figures ($_{0,1,2,3}$) for exponents and subscripts until most of the stencils had been cut. With more time, these and other imperfections could have been adjusted, but in view of the pressure for the appearance of the work, it will not be further delayed.

W. Edwards Deming

TABLE OF CONTENTS

Lectures

Conferences

THE MODERN VIEWPOINT ON THE CLASSICAL THEORY OF PROBABILITY AND ITS APPLICATIONS. TESTS FOR STATISTICAL HYPOTHESES.

Three Lectures delivered at the
Graduate School of the U.S. Department of Agriculture

by
J. Neyman

INTRODUCTION

Since the original titles of my lectures were fixed, I have received a number of letters from the members of the prospective audience, and those letters forced me to modify the original programme.

The conception of probability has been discussed and defined in many different ways, each having its own advantages. 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 so long as the axioms on which it is based are not mutually contradictory and there are no errors in deductions. Among the existing systems of axioms and theories deducible from them we may make a choice. In this we shall naturally be guided by considerations of usefulness or, what frequently amounts to the same, by our personal taste. It is important, however, to make it clear in what theory 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 to others, and which I have always 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 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 during two or three lectures only. 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. Cramer: 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.

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 are what 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 quotations.

In the third and last lecture I shall deal with a 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 that lecture is essentially taken from an article of mine, published in 1929 in the Reports of the First Congress of Slavonic Mathematicians in Warsaw. Its title is "Méthodes nouvelles de vérification des hypothèses statistiques."

LECTURE I: ON THE THEORY OF PROBABILITY

1. Definition of Probability. The probability that I shall define will always relate 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 assuming this nomenclature we may speak also of probabilities of events. These 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 therefore seems to be not less general than the others.

In mathematical definitions the actual wordings used do not matter very much. However they do have some importance, as they may appeal to intuition with different strengths and may differently emphasize the essential source of the concepts introduced. The essential point in the conception of the probability I am going to 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 others that are not. Now all objects A will form what I shall call the fundamental probability set

(F.P.S. for short). This will be denoted also by (A).*

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

Before proceeding any 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 our 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 simply 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 accept the notation $B_4 = B_1B_2$ 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.

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 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

$$P\{B|A\} = k/n = \frac{(B)}{(A)}. \qquad (1)$$

In other words, the probability of an object A having the property

---

* Any letter, e.g. x in parenthesis stands for "all x." This notation is commonly in use.

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 some 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 A may be omitted in the symbol of probability, whereupon $P\{B\}$ will suffice for $P\{B|A\}$. Sometimes, however, we shall have to deal not only with one fundamental probability (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, $P\{B'|AB\}$ may be shortened to $P\{B'|B\}$, and $P\{B'|A\}$ may be shortened to $P\{B'\}$.

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

In order to emphasize this distinction we shall sometimes describe $P\{B'|A\}$ as an absolute probability of B' and the probability $P\{B'|AB\}$ 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 the 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 belongs to each of the elements of the F.P.S. and it (the property E) may be described as the only possible property. It is easily seen that the reverses are true, namely that if E_1 and E_2 are an impossible and an only possible property respectively, then must $P\{E_1\} = 0$ and $P\{E_2\} = 1$. It will be noticed that the relative probability $P\{B'|B\}$ 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 casting 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 casting.

Reading this last sentence once more and comparing it with the definition of probability, (Eq. 1), one will easily see that without further description of the situation, the definition of probability could not be applied to castings. Speaking of "the probability of getting six points on the upper side of a die when casting" and trying to apply the definition of probability we may think of various things.

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

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

(c) Finally we may have in mind some hypothetical series of castings 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 some 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 casting, the last is the most fruitful. We shall see this a little further when I shall speak of the so-called empirical law of big numbers.

Example 3. Consider the familiar expansion $\pi = 3.14159...$ 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 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 one element only and the denominator in the expression (1) serving as the

definition of probability is equal to unity. The numerator may be equal to unity--this if x_{1000} is actually equal to 5--or to zero, if $x_{1000} \neq 5$. 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, the probabilities may refer to some hypothetical probability sets, with assumed properties. This case is the one that the theory is most often concerned with; and it is of extreme importance. Therefore I shall give two illustrations.

Example 4. Consider a set F_1 of n die castings, 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 $P\{six, six|F_2\}$, of a pair of castings 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 $P\{six, six|F_2\}$ consists in the probability $P\{six|F_1\}$ of getting a six in one casting. Assume, for instance, that

$$P\{six|F_1\} = 1/6 \qquad (2)$$

This would mean that among the n castings 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 castings forming F_2 there are exactly

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

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

$$P\{six, six|F_2\} = (n-6)/36(n-1) \qquad (4)$$

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 $P\{six, six|F_2\} = (n-6)/36(n-1)$. Thus, if the probability set F_2 has the properties as specified in the conditions of the problem, then the formula (4) holds good. We may notice at this stage that the properties of a probability set F_2 that are 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 position prevails also 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, consisting of the same number m of elements each, is connected with each of the experiments; and denote by F_i the set corresponding to the ith experiment, i = 1, 2, ..., n. Suppose that whatever be i, the probability of the event E calculated with regard to F_i is the same p, that is,

$$P\{E|F_i\} = p \qquad (5)$$

We may now consider still another probability set, say F_o, the elements of which are all possible combinations of the elements of the sets F_1, F_2, ..., F_n taken n at a time, each element selected from a different set. If each of the sets F_1, F_2, ..., F_n consists of the same number m of elements, then the set F_o will consist of m^n elements.

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

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

But it is important to know what this formula denotes. It is no more and no less than the proportion of elements of the set F_o that have the desired property, consisting of k "events" E and n-k "events" non-E.

As mentioned in this example, again the calculation of the probability $P_{n,k}$ referring to the probability set F_o was based on the probabilities referring to the sets F_1, F_2, ..., F_n and on the structure of the elements of F_o, each of them being composed of those of F_1, F_2, ..., 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_o are some combinations of those of some other sets F_1, F_2, etc., then we shall say that the set F_o is of a higher order than the sets F_1, F_2, ... Thus we may distinguish probability sets of first, second, third, etc. orders.

In example 4 the set F_1 was of first, and the set F_2 of second order. In example 5 the sets F_1, F_2, ... , F_n were of first order and the set F_o 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, if purity of language were demanded, they should really 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 big numbers.

We shall notice now that a description of an experiment as in the examples above amounts really to a description of a probability set. As those were classified, so will be classified the corresponding hypothetical experiments. Therefore we shall speak of experiments of the first, second, third, ... 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 a six, and the letter r for a 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 which are equal to the ordinal numbers of the two elements of F_1 making up one 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 the upper chart on the next page that the number of elements forming F_2 is 66 and that therefore $P\{six, six|F_2\} = 1/66$, 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, ..., 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 being non-E, denoted by G. 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 F_0 by points on a plane, of which the 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 the lower figure on the next page). All the elements of F_0 possess the required property of being composed of

y												
11												ss
10											sr	sr
9										rr	sr	sr
8									rr	rr	sr	sr
7								rr	rr	rr	sr	sr
6							rr	rr	rr	rr	sr	sr
5						rr	rr	rr	rr	rr	sr	sr
4					rr	rr	rr	rr	rr	rr	sr	sr
3				rr	rr	rr	rr	rr	rr	rr	sr	sr
2			rr	rr	rr	rr	rr	rr	rr	rr	sr	sr
1		rr	rr	rr	rr	rr	rr	rr	rr	rr	sr	sr
	1	2	3	4	5	6	7	8	9	10	11	12 x

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

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).

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

$$P\{six, six|F_2\} = 1/66 \quad (Ex.4) \qquad (7)$$

and

$$P\{six, six|F_o\} = 1/36 \quad (Ex.5) \qquad (8)$$

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_o. This emphasizes the fact that the probabilities refer to probability sets and that the failure to specify them properly may, and usually does cause misunderstandings.

2. More general definition of probability. The above definitions and examples are probably 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 all possible. In fact we shall require that the following postulates should 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)), $m(B) > 0$.
(b) If (B) is empty (does not contain any single element) then it is measurable and $m(B) = 0$.
(c) The measure of (A) is greater than zero.
(d) If (B_1), (B_2), ..., (B_n) is any at most denumerable set of measurable subsets, then their sum, (ΣB_i), is also measurable. If the subsets of no two pairs (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 $(\overline{B})$ of objects A not possessing the property B is also measurable and consequently, owing to (d), $m(B) + m(\overline{B}) = m(A)$.

Under the above conditions the probability, $P\{B|A\}$ of an object A having the property B will be defined as the ratio $P\{B|A\} = m(B)/m(A)$.

The probability $P\{B|A\}$, 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 $P\{B_1|B_2\} = m(B_1B_2)/m(B_2)$ 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 $P\{B_1|B_2\} = m'(B_1B_2)/m'(B_2)$. Whatever may be the case we shall have

$$P\{B_1B_2\} = P\{B_1\}\, P\{B_2|B_1\} = P\{B_2\}\, P\{B_1|B_2\} \qquad (9)$$

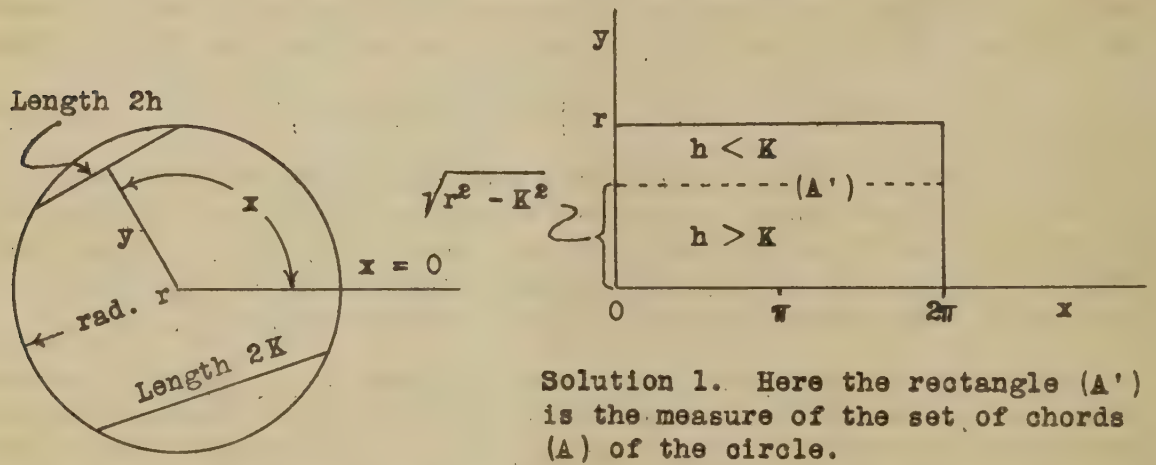
It is easy to see that if the fundamental probability set is finite, then the number of elements in any of its subsets will satisfy the definition of the measure. On the other hand, if (A) is the set of points filling up a certain region in n-dimensional space, then the measure of Lebesgue will satisfy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

If the objects A are not actually points (e.g. if they are certain lines, etc.) the above definitions of probability may be again applied, provided it is possible to establish a one to one correspondence between the objects A and other objects A', forming a class of sets where the measure has already been defined. If (B) is any subset of (A) and (B') the corresponding subset of (A'), then the measure of (B) may be defined as being equal to that of (B'). It is known that a similar definition of measure of subsets of (A) could be done in more than one way. Such is for instance the case when the objects A are the chords in a circle C of radius r and the property B consists of their length 2h exceeding some specified value 2K. It may be useful to consider two of the possible ways of treating this problem (Bertrand's problem).

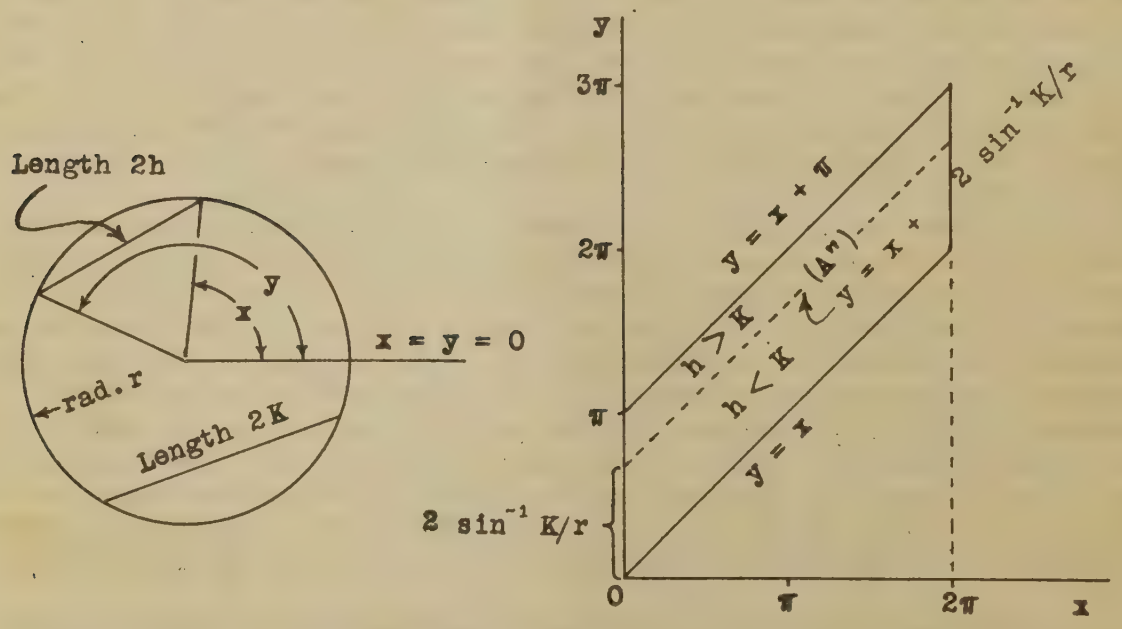
1. Denote by x the angle between any fixed direction and the radius perpendicular to any given chord A, in a circle of radius r. Further, let y be the perpendicular distance of the chord A from the centre of the circle C. Now let A' denote a point on the xy plane with coordinates x and y; then there will be a one to one correspondence between the chords (A) of length $0 \leq 2h \leq 2r$ and the points of a rectangle, say (A'), defined by the inequalities $0 < x \leq 2\pi$ and $0 < y \leq r$. (See the upper figure on the next page). The measure of the set of chords with lengths exceeding $2K < 2r$ could be defined as equal to the area of that part of (A') where $0 \leq y^2 < (r^2 - K^2)$. It follows that the probability in which we are interested is $P\{h > K\} = \sqrt{[1 - (K/r)^2]}$.

2. Denote by x and y the angles between a fixed direction and the radii connecting the ends of any given chord A. If A" denotes a point on a plane with coordinates x and y, then there will be a one to one

correspondence between the chords of the system (A) and the points within the parallelogram (A") determined by the inequalities $0 < x \leq 2\pi$, and $x \leq y \leq x + \pi$ for $0 < x \leq \pi$ and $x \leq y < x + \pi$ for $\pi < x \leq 2\pi$. (See the lower figure of this page). The measure of the set of chords with their lengths exceeding 2K may be defined as being equal to the area of that part of (A") where $x + 2 \text{ arc sin } K/r < y \leq x + \pi$. Starting with this definition, then $P\{h > K\} = 1 - (2/\pi) \text{ arc sin } K/r$.



Solution 1. Here the rectangle (A') is the measure of the set of chords (A) of the circle.



Solution 2. Here the parallelogram (A") is the measure of the set of chords (A) of the circle.

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 is only determined when we define the measure appropriate to the set (A) and its subsets to be considered. We may describe this also differently using the terms "random experiments" and "their results." We may say that to have the 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 leading to the 1st solution (upper figure on the preceding page), we could say that when selecting at random a chord A, we first pick up at random the direction of a radius, all directions being equally probable, and then, equally 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 2d solution (lower figure on the preceding page).

We frequently use this way of speaking, but it is necessary to remember that behind such words, as e.g. "picking up 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 the sentence like the one in quotation marks, just written, is no more than a way of describing the fundamental probability set and the appropriate measure. The conception of "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 up at random a chord, 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 so long as they are clearly defined, our choice must be justified in one way or another. The justification of the way of speaking about the definition of the measure within the fundamental probability set in terms of imaginary random experiments, lies in the empirical fact

which Bortkiewicz* insisted calling the "law of big 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**, by the experiment with throwing a needle** so as to obtain an analogy to the problem of Buffon, and by various sampling experiments based on Tippett's random numbers#.

These examples show that the 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 $\frac{1}{2}$. To get this result we must select what could be called a well balanced coin and we have to work out an appropriate method 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 sufficiently approach, 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.[1]

We shall now draw a few obvious but important conclusions from the definition of the probability 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 the elements of the fundamental probability set (A) possess a certain property B_0, then the absolute probability of B_0,

---

* L. von Bortkiewicz: Die Iterationen, Julius Springer, Berlin, 1917.

** This is mentioned by É. Borel, Éléments de la Théorie des Probabilités. Paris 1910, p.106. I could not find the name of the performer of the experiment.

L. H. C. Tippett: Random Sampling Numbers, Tracts for Computers, No.XV, Edited by Karl Pearson, Cambridge, 1927. Price in England 3/9; in New York, at G. E. Stechert & Co., 31 East 10th Street, $1.25 plus postage.

[1] Cf. some remarks on page 18a.

given any other property B_1, must be equal to unity, so that $P\{B_o|A\} \equiv P\{B_o\} = P\{B_o|B_1\} = 1$. On the other hand, if it is known only that $P\{B_o\} = 1$ then it does not necessarily follow that $P\{B_o|B_1\}$ 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 $P\{a \leq x < b\}$ of x having a value such that $a \leq x < b$.

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 $P\{a \leq x < b|B\} = \frac{1}{2}$. 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 $0 < P\{a \leq x < b|B\} < 1$.

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 $P\{a \leq x < b\}$, and (b) that there are such properties B and such numbers $a < b$ that $0 < P\{a \leq x < b\} < 1$, then x is a random variable in the sense of the above definition.

The probability $P\{a \leq x < b\}$ 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 $P\{a \leq \theta < b|B\}$ may have only values unity or zero according to whether θ falls in between a and b or not.

Keeping in mind the above definitions of the variables in discussing them, we may speak in terms of random experiments. In the sense of the convention adopted above, 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 $\pi = 3.14159\ldots$ 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 λ_{1000} does not depend upon the result of any random experiment.

Frequently we have to consider simultaneously several random variables

$$x_1, x_2, \ldots, x_n \tag{10}$$

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

Denote by E the set of values of the x variables (10). This set could be represented by a point*, to be called the sample point E in an n-dimensional space, say W, the rectangular coordinates of the point being the values $x_1, x_2, \ldots, x_n$. The space W will be called the sample space. Denote by w any region in W and accept the convention that

$$E \in w$$

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

$$P\{E \in w\}$$

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. It must of course be assumed that F(w) is measurable. With that restriction the probability $P\{E \in w\}$ 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 $P\{E \in w\}$ stands for the integral probability law of the variables (10), and if there exists a function p(E) of the x_i such that whatever be w

$$P\{E \in w\} = \iint\ldots\int_w p(E)\, dx_1\, dx_2 \ldots dx_n \tag{11}$$

then the function p(E) is called the elementary probability law of the random variables (10).

It will be noticed that while the integral probability law is a

---

* It is convenient to recall here that mathematicians define a point as a set of numbers.

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.

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

$$\left.\begin{array}{l} x_1,\ x_2,\ \dots,\ x_{n-1},\ x_n \\ \text{and} \\ x_1,\ x_2,\ \dots,\ x_{n-1} \end{array}\right\} \qquad (12)$$

respectively, then

$$p(x_1\ x_2\ \dots\ x_{n-1}) = \int_{-\infty}^{\infty} p(x_1,\ x_2,\ \dots,\ x_{n-1},\ x_n)\ dx_n \qquad (13)$$

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.

2. If there are two sets of n random variables each,

$$x_1,\ x_2,\ \dots,\ x_n \qquad (14)$$

and

$$y_1,\ y_2,\ \dots\ y_n \qquad (15)$$

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

$$\Delta = \frac{d(x_1,\ x_2 \dots\ x_n)}{d(y_1,\ y_2 \dots\ y_n)} \qquad (16)$$

existing and being different from zero almost everywhere and never changing its sign, then the probability laws $p(x_1 \dots\ x_n)$ and $p(y_1 \dots\ y_n)$ of the variables (14) and (15) respectively are connected by the identity

$$p(y_1\ y_2 \dots\ y_n) = p(x_1\ x_2\ \dots\ x_n)\,|\Delta| \qquad (17)$$

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 their simultaneous probability law is known.

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

Consider the set of random variables

$$x_1, x_2, \ldots, x_n \tag{10}$$

Any assumption concerning their probability law (either integral or elementary) is called a statistical hypothesis.

A statistical hypothesis is called <u>simple</u> if it specifies the integral probability law, $P\{E \in w\}$ of the x_i as a single valued function of the region w.

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

The assumption H_1 that

$$p(E) = (1/\sigma\sqrt{2\pi})^n e^{-\Sigma(x_i - \mu)^2/2\sigma^2} \tag{18}$$

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

$$P\{E \in w\} = (1/\sigma\sqrt{2\pi})^n \int\cdots\int_w e^{-\Sigma(x_i - \mu)^2/2\sigma^2}\, dx_1\, dx_2 \ldots dx_n \tag{19}$$

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 the formula (18) 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 (19) we shall be able to calculate the unique numerical value of $P\{E \in w\}$, though sometimes this may be connected with great technical difficulties.

- - - - - - - - - - - - - - - - - - - -

With reference to the discussion on the law of great numbers, particularly in line with certain statements on page 14, it might be useful to remark that for any mathematical theory of probability, which will necessarily always be based on a given F.P.S., it is possible with sufficient care to arrange a set of experiments such that when a long series of them is taken into consideration, the results will approach the theory satisfactorily. For instance, it would be possible to arrange a real laboratory experiment in which chords of a circle are picked up "at random" in such a way that if the performance is repeated many times, the relative frequencies in various classes of lengths of chords actually picked up will approach those of the first solution of Bertrand's problem (page 11 and the upper figure of page 12). It would also be possible to arrange another experiment in which the frequencies in the same class intervals approach those of the second solution (page 11 and the lower figure of page 12). To assume, without actual experience and comparison (see also Lecture II, particularly page 22), that a series of real experiments conforms in any degree to the theory of probability based on a particular F.P.S. is presumptuous; and failure to recognize this point has more than once brought grave difficulties and inconsistencies. Editor.

- - - - - - - - - - - - - - - - - - - -

An article that should be mentioned in connection with Lectures I and II is one by D. J. Struik, "On the foundations of the theory of probabilities" Philosophy of Science 1, 50-70, 1934 (letter from Dr. Neyman dated 7th January 1938).

LECTURE II: ON PROBABILITY AND EXPERIMENTATION

1. Abstract character of mathematical theories and possibilities of applications. It is probable that many listeners at my first lecture were disappointed. They are engaged in various applications of probability to practical problems, and such problems must be the only cause of their interest in the theory of probability. They may feel that they have no use for a theory in which "experiments," "results," etc., everything that is of utmost importance to them, are treated only as picturesque descriptions of probability sets and measures. Those may be good for mathematicians, they would 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 it is at all possible to satisfy that part of the 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 conceptions 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 conceptions 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 this table is a plane. But if we look at this surface through a good magnifying glass we shall immediately say that it is certainly not plane. If we say that it is, we mean that for practical purposes it could be considered as 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 conceptions and some realities; we are satisfied at finding 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 some 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 implements you desire and measure the sides of all the triangles as actually drawn. Using those 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 this is due to the errors of measurement. That 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 whether 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 conceptions not existing in the real world. But there are real objects that correspond to these abstract conceptions in a certain sense, and numerical values of mathematical formulas more or less agree with the results of actual measurements. In earlier stages of any branch of mathematically treated natural science we are satisfied with only 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 how the mathematical theory of probability can be usefully applied in natural science.

2. Random experiments and the empirical law of big 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 conceptions and not with real objects of any kind. Therefore the application of such a theory will be possible only if there can be established a bridge or a correspondence between the conceptions of the theory and the real facts. The actual applications must be preceded by numerous checks and rechecks of the permanency and the accuracy of such correspondence. If this is judged sufficiently accurate and found sufficiently permanent, then the predictions--the final aim of any science--based on the mathematical theory of probability, will have some views of success. Otherwise the theory may be interesting by itself, but useless from the point of view of application.

What is then the category of facts that correspond to conceptions of the theory of probability as described in my first lecture? What is the meaning of that correspondence?

The category of these facts may be described as the results of random experiments. It is impossible to give an exact definition of experiments that are called random. Equally it would be impossible to give the definition of such objects in the real world deserving the description "plane", "straight line", etc. Instead of speaking of real objects we shall speak of abstract conceptions. At most we can give a rough description illustrating it with some examples 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 utmost care to keep the 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 electrical motor so as to impart a constant initial velocity to the coin. The experiments may be carried on in a closed room with no noticable air currents; the coin may be put into the machine always in the same way; and then--I am practically certain that the results of the repeated experiments will vary. Perhaps very 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 the roulette. A well constructed roulette with an electrically regulated start will yield varying results.

(c) Those were types of random experiments arranged by men. 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 observations being considered as a single experiment. It will be found that however constant be the conditions of the consecutive experiments, their 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.

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

Let N be a fairly large number, say 1000 or so, and n any other positive integer. 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 shall obtain the ratio

$$f = M/Nn . \qquad (1)$$

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, ..., 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

$$F_{n,k} = m_k/N \qquad (2)$$

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 in as uniform conditions as possible, and the number N is large, then the relative frequency $F_{n,k}$ appears to be very nearly equal to the familiar formula

$$\frac{n!}{(n-k)!\,k!}\,(1-f)^{n-k}\,f^{k} \qquad (3)$$

In other words, the relative frequency $F_{n,k}$ relating to a series of second order experiments is connected with the relative frequency f of the first order experiments very nearly in the same way as the probability $P_{n,k}$ discussed in my first lecture (page 7) and relating to the second order probability set, 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 castings 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 considerably differ from 1/6, depending on the structure of the die and the exact conditions of casting. 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

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

The above fact, which has been found empirically* many times, could be described in a more general way by saying that usually the single random experiments and various groups of these experiments behave as if they tended to reproduce certain first order probability sets, corresponding to first order trials, and an appropriate second order probability set. This fact may be called the empirical law of big numbers. I want to emphasize that this law applies not only to the simple case discussed above, connected with the binomial formula, but seems to be perfectly general, in the same sense in which we use the word general with respect to any other "general law" observed in the outside world. Whenever it fails, we explain it 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 big numbers. Then, as it is our custom to do, we expect them to behave similarly in the future, and 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) Wishing to treat certain phenomena by means of the theory of probability we must find some element of those phenomena that could be considered as random, following the law of big 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.

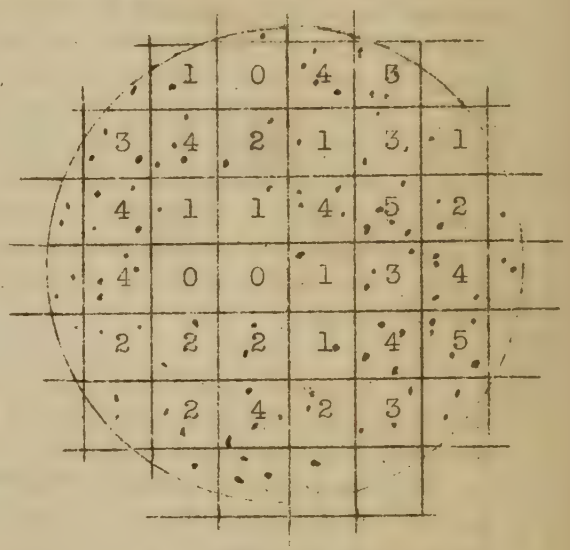
(iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future.

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

---

* See for example L.von Bortkiewicz: Die Iterationen
Julius Springer, Berlin, 1917.

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 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 also 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 one particular living bacterium B contained in the liquid that was poured on to the plate. All the operations performed with the liquid and the plate resulting in fixing the bacterium B in some point are considered in the mathematical model 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. Those 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 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



$$P_k = e^{-\lambda} \lambda^k / k! \qquad (5)$$

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

$$P\{a \leq k < b\} = \sum_{k=a}^{b} e^{-\lambda} \lambda^k / k! \text{ for } 0 \leq a < b \qquad (6)$$

If this mathematical model could be assumed to correspond accurately to the actual experiments in the sense of the word as 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 being produced by two or more unconnected bacteria.

In order to solve the question whether the number k of colonies within a square could be considered as a random variable and whether its probability law could be represented by the formula (5) my colleagues have performed a series of experiments summarized in the following table.

Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten.
II. Abteilung. 1936, Bd. 95

Comparison of distributions of colonies with Poisson Law .

T. Matuszewski, J. Supinska und J. Neyman

k	Agarplatten										k
	1 Schizosacch. Pombe		2 Strept. lactis		3 Thermob. helvetic.		4 Bact. fluorescens		5 Hefe u. Bakterien		
	m′	m	m′	m	m′	m	m′	m	m′	m	
0	5	6,1	26	27,5	59	55,6	83	75,0	0 } 5	4,6	0
1	19	18,0	40	42,2	86	82,2	134	144,5	5 }		1
2	26	26,7	38	32,5	49	60,8	135	139,4	9	11,0	2
3	26	26,4	17	16,7	30	30,0	101	89,7	23	20,9	3
4	21	19,6	5 } 7	9,1	15 }		40	43,3	33	29,6	4
5	13	11,7	2 }		3 } 20	15,4	16	16,7	32	34,0	5
6	4 }				2 }		3 }		32	31,8	6
7	3 } 8	9,5					2 } 7	7,4	24	25,8	7
8	1 }						2 }		13	18,3	8
9									12	11,6	9
10									8	6,7	10
11									7 } 9	5,7	11
12									2 }		12

k	Gelatineplatten										k
	6 Sacch. cerevisiae		7 Bact. Beijerincki		8 Bact. Maerck.		9 Bact. rancens		10 Bewegliche Stäbchen		
	m′	m	m′	m	m′	m	m′	m	m′	m	
0	8	6,8	0 } 12	10,3	7	3,9	3	2,1	60	62,6	0
1	16	16,2	12 }		11	10,4	7	8,2	80	75,8	1
2	18	19,2	18	16,7	11	13,7	14	15,8	45	45,8	2
3	15	15,1	13	22,4	11	12,0	21	20,2	16	18,5	3
4	9	9,0	27	22,7	7	7,9	20	19,5	8 } 9	7,3	4
5	4 }		19	18,3	3 }		19	15,0	1 }		5
6	2 } 7	6,7	16	12,3	2 }		7	9,6			6
7	0 }		6 }		1 } 8	7,1	6 }				7
8	1 }		4 } 11	13,3	1 }		1 } 9	9,6			8
9			1 }		1 }		0 }				9
10							2 }				10
Platten-Nr.	1		2	3	4	5	6	7	8	9	10
χ^2	0,77		1,61	4,05	3,47	4,94	0,30	6,67	3,21	2,63	1,09
P	0,97		0,66	0,26	0,63	0,84	0,97	0,25	0,53	0,85	0,78

The values of k are the numbers of colonies within the squares into which the whole plate was divided. m' and m denote the observed and the expected numbers of squares having the number k of colonies. The kind of bacteria analyzed is stated at the top of each pair of columns. 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 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 carried out up to the present 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 the conclusions concerning those 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 arriving 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. The table following is reproduced from an article in Biometrika, and represents a comparison between the Poisson law, Eq. 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 altogether 12 comparisons, of which 11 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 not as satisfactory as in the preceding example. It seems to follow that if the samples of clover seed are drawn by the method employed by Schindler, then the conclusions concerning them drawn from the mathematical model involving the Poisson law, Eq. 5, 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 the two authors quoted--that the number of dodders growing per square could rightly be considered as a random variable following the Poisson law.

Example 3. As mentioned above, if certain experiments show definite divergence from a mathematical model that is strongly suggested by intuition, then the divergence may be ascribed to "errors of experimentation," and efforts may be made to change the experimental technique with the hope that it may result in a more satisfactory agreement between experimental data and the theory.

J. PRZYBOROWSKI AND H. WILENSKI

Biometrika 27, 277, 1935

Distribution of Dodder in Samples of Clover Seed. (Schindler's Experiments.)

1			2			3			4		
k	N_k	$N.P_k$	k	N_k	$N.P_k$	k	N_k	$N.P_k$	k	N_k	$N.P_k$
0	168	183.94	0	599	606.53	0	382	389.40	0	284	303.27
1	205	183.94	1	315	303.27	1	111	97.35	1	170	151.63
2	94	91.97	2	74	75.82	2	7	12.17 } 13.25	2	39	37.91
3	26	30.66	3	12	12.64 } 14.38	over 2	0	1.08 }	3	7	6.32 } 7.19
4	6	7.66 }	over 3	0	1.74 }				over 3	0	0.87 }
5	1	1.53 } 9.49									
over 5	0	0.30 }									
m		1	m		0.5	m		0.25	m		0.5
n'		5	n'		4	n'		3	n'		4
χ^2		5.1990	χ^2		0.9848	χ^2		5.0027	χ^2		3.4863
P		0.160	P		0.600	P		.000	P		0.180

5			6			7			8		
k	N_k	$N.P_k$	k	N_k	$N.P_k$	k	N_k	$N.P_k$	k	N_k	$N.P_k$
0	795	774.64	0	447	452.42	0	473	475.61	0	295	303.27
1	94	116.20	1	51	45.24	1	26	23.78	1	153	151.63
2	11	8.71 } 9.20	2	2	2.26 } 2.34	over 1	1	0.61	2	44	37.91
over 2	0	0.45 }	over 2	0	0.08 }				3	8	6.32 } 7.19
									over 3	0	0.87 }
m		0.15	m		0.1	m		0.05	m		0.5
n'		3	n'		3	n'		3	n'		4
χ^2		5.1286	χ^2		0.8477	χ^2		0.4709	χ^2		1.3075
P		.000	P		.198	P		0.319	P		0.533

9			10			11		
k	N_k	$N.P_k$	k	N_k	$N.P_k$	k	N_k	$N.P_k$
0	22	16.42	0	0	1.08	0	0	0.09
1	29	41.04	1	3	5.39	1	0	0.66
2	55	51.30	2	13	13.48	2	1	2.49
3	43	42.75	3	15	22.46	3	4	6.22
4	34	26.72	4	33	28.07	4	9	11.67
5	10	13.36	5	28	28.07	5	16	17.50
6	3	5.57 }	6	24	23.40	6	19	21.87
7	4	1.99 } 8.41	7	21	16.71	7	19	23.44
over 7	0	0.85 }	8	10	10.44	8	26	21.97
			9	8	5.80 }	9	19	18.31
			10	0	2.90 } 10.90	10	15	13.73
			11	4	1.32 }	11	14	9.36
			over 11	1	0.88 }	12	5	5.85 }
						13	6	3.38 }
						14	3	1.81 } 12.68
						15	3	0.90 }
						over 15	1	0.74 }
m		2.5	m		5	m		7.5
n'		7	n'		10	n'		13
χ^2		8.7617	χ^2		7.0406	χ^2		9.8060
P		0.120	P		0.532	P		0.548

Authors' own experiment with known

$\lambda = 2$

k	N_k	$N.P_k$
0	56	67.67
1	156	135.34
2	132	135.34
3	92	90.22
4	37	45.11
5	22	18.04
6	4	6.02 }
7	0	1.72 } 8.29
8	1	0.43 }
over 8	0	0.12 }

$n' = 7$, $n = n' - 1$,
$\chi^2 = 8.9169$, $P = 0.179$

k = number of dodder seeds in a sample.
N_k = observed frequency.
$N.P_k$ = expected frequency.

n' = number of groups.
n = number of degrees of freedom for the chi-square tables.

Another way, of course, having the same aim, is to alter the mathematical model. However, frequently the first method is more satisfactory. Examples of such attempts to bring particular experiments into an agreement with probability models are constantly carried out in big manufacturing works, arriving at what is called "statistically controlled production." In this respect the reader will find it interesting to consult the book by Dr. W. A. Shewhart, of the Bell Telephone Laboratories, The Economic Control of Quality, Van Nostrand, New York, 1931.

I shall give here a description of the efforts of this kind which seem to be particularly interesting.

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, three, four, sometimes five, portions, and those are separately analyzed. 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.

Owing to 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

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad (7)$$

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. It is obviously useless to try to obtain an accurate 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 of 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*

$$\left.\begin{aligned} f_1 &= \bar{x} - t_\alpha s'/\sqrt{n} \\ \text{and} \quad f_2 &= \bar{x} + t_\alpha s'/\sqrt{n} \end{aligned}\right\} \qquad (8)$$

have a remarkable property. Here $\bar{x}$ is the arithmetic mean of the measurements x_i, n their number, s' their estimated standard deviation**, and t_α the value of Fisher's t corresponding to the number of degrees of freedom

---

* J. Neyman, "Outline of a theory of estimation," Phil. Trans. Royal Soc. A236, 333-380, 1937. See also the conferences on estimation and confidence intervals, pp.127-142, and 143-160 respectively.

** That is, s' is an estimate of σ; $s'^2 = \sum (x_i - \bar{x})^2/(n-1)$; see p.135.

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 (7), 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, discovered about 1930,* permits the estimation of μ in a form of a random experiment. We perform the experimental analysis, obtaining the values of the x_i, and then state that

$$\bar{x} - t_\alpha s'/\sqrt{n} \leq \mu \leq \bar{x} + t_\alpha s'/\sqrt{n} \qquad (9)$$

We may be wrong in this statement, but if the x_i do follow the law (7), the probability of our being correct is equal to $\alpha = .99$; 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

$$f_2 - f_1 = 2t_\alpha s/\sqrt{n} \qquad (10)$$

is unsatisfactory.

In what preceded, the value of σ in Eq.7 was considered unknown. If, however, σ is known, then the confidence interval will be written

$$\bar{x} - T_\alpha \sigma/\sqrt{n} \leq \mu \leq \bar{x} + T_\alpha \sigma/\sqrt{n} \qquad (11)$$

where T_α is the value of t_α corresponding to an infinite number of degrees of freedom in the estimate of σ. What it 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 then the values of t_α are respectively

$$t_{.01} = 63.657 \text{ if } n = 2$$
$$t_{.01} = 9.925 \text{ if } n = 3$$
$$t_{.01} = 5.841 \text{ if } n = 4$$
$$\text{etc.}$$

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 (7), but also that the standard deviation σ has permanently this or that particular numerical value, then the same few parallel analyses could be used to provide 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, it must obviously be interested (i) in keeping the value of σ constant over long periods of time; and (ii) in estimating

---

* See references on pages 157 and 158.

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 the formula (11) instead of (9) we must see whether the measurements that are being obtained do agree with the mathematical model involving a constant σ. Otherwise a repeated application of formula (11) 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 this table, s_i^2 is the estimated variance of four parallel analyses, and s_o^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 was actually constant during such a period, then the value of s_o^2 would be its very accurate estimate and the mathematical model adopted would imply a known distribution of the ratio $v = s_i^2/s_o^2$.

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 its confidence intervals based on formula (11), 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.

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

Przyborowski, Polish Agric. Forest. Journ. vol. 30, 1933.

1925			1927			1928		
$v = s_i^2/s_o^2$	Obs.	Exp.	$v = s_i^2/s_o^2$	Obs.	Exp.	$v = s_i^2/s_o^2$	Obs.	Exp.
0 - 1	76	35.6	0 - 1	44	24.8	0 - 1	49	31.0
1 - 2	30	41.0	1 - 2	27	28.6	1 - 2	31	35.7
2 - 3	20	32.4	2 - 3	13	22.6	2 - 3	27	28.2
3 - 4	14	23.3	3 - 4	19	16.3	3 - 4	18	20.3
4 - 5	10	16.1	4 - 5	7	11.2	4 - 5	9	14.0
5 - 6	4	10.8	Above 5	15	21.5	5 - 6	5	9.4
Above 6	25	20.0				Above 6	17	17.4
$\chi^2 = 65.101$ $P = .00000$			$\chi^2 = 22.928$ $P = 0.0Q0127$			$\chi^2 = 15.217$ $P = 0.0094$		

1929			1930			1931			1929 - 1931		
$= s_i^2/s_o^2$	Obs.	Exp.	$v = s_i^2/s_o^2$	Obs.	Exp.	$v = s_i^2/s_o^2$	Obs.	Exp.	$v = s_i^2/s_o^2$	Obs.	Exp.
0 - 1	35	30.8	0 - 1	25	27.0	0 - 1	20	20.7	0 - 1	80	78.5
1 - 2	27	35.5	1 - 2	33	31.1	1 - 2	20	23.8	1 - 2	81	90.4
2 - 3	30	28.0	2 - 3	27	24.6	2 - 3	15	18.8	2 - 3	71	71.4
3 - 4	19	20.2	3 - 4	20	17.7	3 - 4	22	13.5	3 - 4	61	51.4
4 - 5	13	13.9	4 - 5	13	12.2	4 - 5	15	9.3	4 - 5	40	35.4
5 - 6	13	9.3	Above 5	18	23.4	Above 5	12	17.9	5 - 6	24	23.8
Above 6	18	17.3							6 - 7	14	15.7
									7 - 8	13	10.2
									Above 8	11	18.2
$\chi^2 = 4.332$ $P = 0.36$			$\chi^2 = 2.084$ $P = 0.72$			$\chi^2 = 12.068$ $P = 0.017$			$\chi^2 = 7.157$ $P = 0.41$		

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

(1) a mathematical theory, e.g., the one described in my first lecture;

(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 either with (2) or (3), 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 the random experiments here falling under (2). The theory presented in my first lecture is one of the types coming under this heading. The theory of Richard von Mises is another. (ii) when 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 up by Harold Jeffreys.** It will be noticed that the theory of probability of my first lecture has nothing to do with the "state of mind," though having found that the probability of a certain property is equal e.g. to 0.0001, the state of our mind will probably 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 a precipice. 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 empirically checking whether the correspondence is satisfactory.

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

---

* Compare with H. Levy and L. Roth: Elements of Probability, Oxford, 1936, p.15.

** See Jeffrey's Scientific Inference, Cambridge, 1931, and numerous papers in the Proceedings of the Royal Society (series A) and in the Proceedings of the Cambridge Philosophical Society.

LECTURE III: ON THE TESTING 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 I shall use the conceptions discussed in my first two lectures and perhaps some more. It would be impossible to give all the necessary definitions here and I must assume them to be known.

The traditional procedure in testing statistical hypotheses is commonly known, but being traditional, the opinions concerning its exact nature vary. I shall describe here one of the versions 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.

Having to test some specified (in early stages, very vaguely specified) statistical hypotheses H concerning the random variables

$$x_1, x_2, \ldots, x_n$$

we used to choose some function T of those x_i 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 a difficult one, consisted in deducing an accurate or at least an approximate probability law $p(T|H)$, 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 off 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 $p(T|H)$ is small compared with its maximum, then the sample E would be called improbable, or the hypothesis H improbable, and inversely. You will certainly remember instances where both very small and very large values of chi-square are supposed to suggest that something is wrong.

Having obtained an improbable sample in the above sense 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 observe 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 a first systematically applied test, the Pearson chi-square of 1900, and has worked, on the whole, satisfactorily. However, now we have become sophisticated and desire to have a theory of tests. Before all we want to know why should we use this or that particular

function T of the x_i as a criterion? Why should we test the goodness of fit by calculating

$$\chi^2 = \Sigma (m - m')^2/m \qquad (1)$$

and not, say

$$\chi'^2 = \Sigma (m - m')^2/m' \qquad (2)$$

or

$$\chi''^2 = \Sigma |m - m'| /m \qquad (3)$$

or anything else? What is the actual meaning of a statistical test? What is the principle of choosing between several tests that may be suggested for the same hypothesis? It is the purpose of the present lecture to discuss some of these questions and to explain some 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 some particular hypothesis H, should we consider that hypothesis only, or something else? It is known that some statisticians are of the opinion that good tests could be devised on consideration of the hypothesis tested only. My opinion is that this is impossible and that if satisfactory tests are actually devised without explicit consideration of something 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. It is rather difficult to discuss what an author may have in his mind subconsciously, or even consciously. But it is easier to consider situations that may present themselves when we are forced to select a test for a particular hypothesis H with nothing to base our device on except this 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 could not think of all the suggestions that could be 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 closer 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 the above 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 using principle I only 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.

Having in view the above situation 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 having in view the testing of 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 also do not go beyond the consideration of the hypothesis tested. For example we may require some particular properties of the functions T to be used as criteria, e.g. that they should be symmetrical with respect to the random variables, etc. However I could not think of any such limitation that would seem reasonable. There are particular cases known when a recognized criterion is not symmetrical with respect to all the x_i, for instance when it is represented either by the smallest or by the greatest of all the observations. Therefore, without claiming that the two propositions which I am going to prove below provide decisive evidence that it is absolutely impossible to base the 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 the argument like "use T_1 rather than T_2 because its standard error is smaller" is not by itself persuasive. Let us now go into some details.

2. Insufficiency of Principle I. Consider the system of n random variables

$$x_1, x_2, \ldots, x_n$$

known to be independent and following the normal law

$$p(x_1 \ldots x_n) = (1/\sigma\sqrt{2\pi})^n \, e^{-\Sigma (x_i - \mu)^2/2\sigma^2} \qquad (4)$$

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 that invented by Student, namely, to calculate

$$z = \bar{x}/s \qquad (5)$$

where

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

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

$$p(z) = C(1 + z^2)^{-\frac{1}{2}n} \qquad (7)$$

where

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

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

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

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 is to have the following properties:

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

$$p(\zeta) = C(1 + \zeta^2)^{-\frac{1}{2}n} \qquad (10)$$

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

$$|z\zeta| \leq 1 \qquad (11)$$

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. Owing to (11), whenever $|\zeta|$ is large then $|z|$ must be small and inversely, and it follows that 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. Therefore, whenever one of the criteria has a large absolute value we are compelled to choose the one the verdict of which 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

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

and

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

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 what I shall call Proposition a, the ratio

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

has the properties 1 and 2 as described above.

In order to prove 1, it will be 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

$$p(x_1 \ldots x_n) = [1/\sigma\sqrt{(2\pi)}]^n e^{-\Sigma x_i^2/2\sigma^2} \tag{15}$$

Let us introduce a new system of random variables

$$y_1, y_2, \ldots, y_n$$

connected with the x_i by the following formulas

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

It will be noticed that

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

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

$$y_2 = (x_1 + x_2)/\sqrt{2} \tag{18}$$

and having regard to (13) we shall have

$$s'^2 = (1/n)\,(y_2^2 + y_3^2 + \ldots + y_n^2) \tag{19}$$

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

$$p(y_1\ y_2 \ldots y_n) = p(x_1\ x_2 \ldots x_n)\ |\Delta| \quad \text{(Eq. 17, page 17)} \quad (20)$$

where $|\Delta|$ is the Jacobian defined by Eq. (16) of page 17, and the x_i in the right-hand side should be expressed in terms of the y_i. Easy calculations give

$$p(y_1\ y_2 \ldots y_n) \equiv p(\bar{x}'\ y_2 \ldots y_n) = [\sqrt{n}/\sigma\sqrt{(2\pi)}]^n e^{-n(\bar{x}'^2 + s'^2)/2\sigma^2} \quad \ldots\ (21)$$

where s'^2 stands for the sum of squares (19). Our next step must consist in introducing still another system of variables

$$u_1, u_2, \ldots, u_n$$

one of which would be identical with $\bar{x}'$ and another with s'. We shall put

$$\left.\begin{aligned} \bar{x}' &= u_1 \\ y_2 &= \sqrt{n}\, u_2 \cos u_n \cos u_{n-1} \ldots \cos u_4 \cos u_3 \\ y_3 &= \sqrt{n}\, u_2 \cos u_n \cos u_{n-1} \ldots \cos u_4 \sin u_3 \\ y_4 &= \sqrt{n}\, u_2 \cos u_n \cos u_{n-1} \ldots \sin u_4 \\ &\vdots \\ y_n &= \sqrt{n}\, u_2 \sin u_n \end{aligned}\right\} \quad (22)$$

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

$$\left.\begin{aligned} -\infty < u_1 &< +\infty \\ 0 < u_2 & \\ 0 \le u_3 &< 2\pi \\ -\tfrac{1}{2}\pi < u_i &< +\tfrac{1}{2}\pi \quad i = 4, 5, \ldots, n \end{aligned}\right\} \quad (23)$$

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

It will be easily seen that

...

$$u_2^2 = (1/n)\,(y_2^2 + y_3^2 + \ldots + y_n^2) \qquad (24)$$

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

$$\left| \frac{d(\bar{x}', y_2, \ldots, y_n)}{d(u_1, u_2, \ldots, u_n)} \right| = u_2^{n-2} \cos u_4 \cos^2 u_5 \cos^3 u_6 \ldots \cos^{n-3} u_n \qquad (25)$$

and it follows that

$$p(u_1\, u_2 \ldots u_n) = [\sqrt{n}/\sigma\sqrt{(2\pi)}]^n\, u_2^{n-2}\, e^{-n(u_1^2 + u_2^2)/2\sigma^2} \cdot \cos u_4 \cos^2 u_5 \ldots \cos^{n-3} u_n \qquad (26)$$

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 $u_3, u_4, \ldots, u_n$ from $-\infty$ to $+\infty$. Owing to the fact that the integrand differs from zero only within the limits shown in (23), and that these limits for $u_3, u_4, \ldots, u_n$ do not depend on the values of u_1 and u_2, we shall have at once

$$p(u_1\, u_2) = C_1[\sqrt{n}/\sigma\sqrt{(2\pi)}]^n\, u_2^{n-2}\, e^{-n(u_1^2 + u_2^2)/2\sigma^2} \qquad (27)$$

wherein

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

and the region of integration, w, is determined by

$$\left.\begin{aligned} 0 \le u_3 &\le 2\pi \\ -\tfrac{1}{2}\pi < u_i &< +\tfrac{1}{2}\pi \quad \text{for } i = 4, 5, \ldots, n \end{aligned}\right\} \qquad (29)$$

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

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

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 Eq. (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 $|z\zeta| \le 1$. For this purpose notice that, whatever the real numbers a and b, we shall have

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

and therefore

$$2\,|ab| \leq a^2 + b^2 \tag{32}$$

It follows further that for any real numbers a and b,

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

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

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

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, owing to (33)

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

Comparing (34) and (36) we see that

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

an inequality between the squares of the quasi mean and 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) must

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

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

$$(\bar{x}^2/s^2)\,(\bar{x}'^2/s'^2) \leq 1 \tag{41}$$

---

* The sign Σ, unless accompanied by other indications, will signify summation over i from 1 to n; i.e. i = 1, 2, ..., n.

which is equivalent to $|z\zeta| \leq 1$, or Eq.(11) of page 36. This fulfills the proof of the part 2 of proposition a, showing that the principle I by itself is not sufficient for a choice between alternative criteria that may be suggested for treating a given hypothesis.

3. Consequences of supplementing Principle I by Principle II. We shall now show that Principle I could not be usefully 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_o, all following the same probability law as determined by H_o, 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_o. We have already seen that if the "Student's hypothesis" (page 36) 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 the sample E' observed in some particular case, not all the x_i being equal to zero, it is possible to find a criterion, say ζ°, 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°; and this will lead to its rejection. Thus in all cases, the only exception being when all the 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

$$E' \equiv x'_1, x'_2, \ldots, x'_n \tag{42}$$

observed in some particular case, one at least of the x_i being different from zero, it is possible to define a criterion ζ° represented by a function of the x_i and having the following properties:

(i) The probability law of ζ°, as determined by H, is the same as that of Student's z and ζ, Eq.(7) page 36.

(ii) The value $\zeta^\circ(E')$ of ζ°, calculated for the sample E', is equal to infinity.

It will be noticed that ζ° will have to be adjusted to the sample E' already observed. Therefore the values (42) will have to enter into the expression of ζ°. They are constant numbers and will play the rôle of coefficients. On the other hand, ζ° 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 $x'_1, x'_2, \ldots, x'_m$

are different from zero, $m \leq n$. Before defining ζ° we shall need the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$, which are to be connected with the x'_i by the n equations

$$\alpha_i = x'_i / \sqrt{(x'^2_1 + x'^2_2 + \ldots + x'^2_n)}, \quad i = 1, 2, \ldots, n \qquad (43)$$

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

$$\Sigma\, \alpha_i^2 = 1 \qquad (44)$$

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

$$\zeta^\circ = \bar{x}'' / s'' \qquad (45)$$

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

$$\bar{x}'' = (\alpha_1 x_1 + \ldots + \alpha_n x_n) / \sqrt{n} \qquad (46)$$

and

$$s''^2 = (1/n) \Sigma\, x_i^2 - \bar{x}''^2 \qquad (47)$$

It will be noticed that if $\alpha_i = 1/\sqrt{n}$ for $i = 1, 2, \ldots, n$, 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

$$v_1, v_2, \ldots, v_n$$

the elementary probability law of which (Lecture I, page 16) as determined by H is

$$p(v_1\ v_2\ \ldots\ v_n) = [\sqrt{n}/\sigma\sqrt{(2\pi)}]^n e^{-n(v_1^2 + s''^2)/2\sigma^2} \qquad (48)$$

wherein

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

To show that $v_1, \ldots, v_n$ exist with the probability law (48) we may introduce

$$\beta_k = \alpha_k [(\alpha_1^2 + \ldots + \alpha_{k-1}^2)(\alpha_1^2 + \ldots + \alpha_k^2)]^{-\frac{1}{2}} \text{ for } k = 2, 3, \ldots, m \leq n \qquad (50)$$

and relate $v_1, \ldots, v_n$ to $x_1, \ldots, x_n$ by the following systems of equations:

$$\left.\begin{aligned}
x_1 &= \sqrt{n}.\ \alpha_1 v_1 + \alpha_1 (\beta_2 v_2 + \beta_3 v_3 + \beta_4 v_4 + \ldots + \beta_m v_m) \\
x_2 &= \sqrt{n}.\ \alpha_2 v_1 - (\alpha_1^2/\alpha_2)\, \beta_2 v_2 + \alpha_2 (\beta_3 v_3 + \beta_4 v_4 + \ldots + \beta_m v_m) \\
x_3 &= \sqrt{n}.\ \alpha_3 v_1 - [(\alpha_1^2 + \alpha_2^2)/\alpha_3]\, \beta_3 v_3 + \alpha_3 (\beta_4 v_4 + \ldots + \beta_m v_m) \\
&\ \ \vdots \\
x_k &= \sqrt{n}.\ \alpha_k v_1 - [(\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_{k-1}^2)/\alpha_k]\, \beta_k v_k \\
&\qquad + \alpha_k (\beta_{k+1} v_{k+1} + \ldots + \beta_m v_m)
\end{aligned}\right\} \quad (51)$$

$k = 2, 3, \ldots, m-1$. Finally, if $m = n$, then

$$x_n = \sqrt{n}.\ \alpha_n v_1 - [(\alpha_1^2 + \ldots + \alpha_{n-1}^2)/\alpha_n]\, \beta_n v_n \qquad (52)$$

Otherwise,

$$x_i = v_i \quad \text{for } i = m+1, \ldots, n \qquad (53)$$

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

$$\begin{aligned} v_1 &= (1/\sqrt{n})\ (\alpha_1 x_1 + \ldots + \alpha_n x_n) \\ &\equiv \bar{x}'' \end{aligned} \qquad (54)$$

and that

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

The Jacobian $|\Delta| \equiv \dfrac{d(x_1, x_2, \ldots, x_n)}{d(v_1, v_2, \ldots, v_n)} = \sqrt{n}$, as is not difficult to work out from Eqs. (51),(52),(53). From Eq. (55), and the value of the Jacobian, it follows by applying Eq. (17) of page 17 (Lecture I) that if Eq. (15) is the simultaneous elementary probability law of $x_1\ x_2 \ldots x_n$, then that of $v_1\ v_2 \ldots v_n$ must be as written in Eq. (48).

Now Eq. (48), being the same form as Eq. (21), and formulas (49) like (17) and (19), it is clear that the steps required to deduce $p(\zeta^\circ)$ from (48) would be identically those already taken to deduce $p(\zeta)$ from (21), and the result must be the property (i) for ζ°; so the proof is completed.

Proof of part (ii) of proposition b. We must now prove the other statement (ii) on page 41 concerning ζ°; namely, we must prove that if in the expression for ζ° we substitute, instead of the random variables x_i, the particular observed values x'_i of (42) in terms of which the function ζ° has been defined, then the value $\zeta^\circ(E')$ of ζ° will be found equal to infinity. Replacing x_i by x'_i in Eq.(46), and remembering that the coefficients α_i therein have already been defined by Eq.(43) in terms of x'_i, we easily find that the value of the pseudo mean calculated for the sample E' is

$$x''(E') = \sqrt{[(x'_1{}^2 + x'_2{}^2 + \ldots + x'_n{}^2)/n]} > 0 \qquad (56)$$

since at least one of the numbers x'_i is different from zero. Further, substituting x'_i for x_i in Eq.(47) to calculate the pseudo S.D. $s''(E')$, we find it to be zero. It follows from Eq.(45) that $\zeta^\circ(E') = x''(E')/s''(E') = \infty$, and this completes the proof of part (ii) of proposition b.

For the one particular sample E' already drawn, ζ° 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 be generally applied for making a choice from 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 reasonings could be generalized, but such generalization would not produce any new really 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 some particular statistical hypothesis H_o, 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 whenever we attempt to test a hypothesis we do admit, subconsciously perhaps, that there are hypotheses that are contradictory or, as we call them, alternative, to the one that is 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 it.

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_o. It is easy to see that there are two kinds:

> After having applied a test we may decide to reject the hypothesis H_o, when in fact, though we do not know it, it is actually true. This is called an <u>error of the first kind</u>.
>
> After having applied a test we may decide not to reject H_o (this may be described for short by saying that we "accept H_o") when in fact H_o is wrong, and therefore some alternative hypothesis H' is true. This is called an <u>error of the second kind</u>.

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

It is nothing but a rule according to which we sometimes reject the hypothesis tested and sometimes accept it (in the sense of the word explained above), according to whether the observations available possess some properties specified by the rule. The observations are some n numbers

$$x_1, x_2, \ldots, x_n .$$

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, are to lead to a rejection of H. These positions usually fill up a certain region, w, which is called the <u>critical region</u> or the <u>region of rejection</u>.

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

Let us illustrate this by one 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,... n. Denote by M the total number of observations and by m_i the number of

---

* It may be helpful here to refer back to Lecture I, page 16.

them belonging to the ith category.

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

$$\chi^2 = \Sigma (m_i - Mp_i)^2 / Mp_i \tag{57}$$

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 the 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 $m_1, m_2, \ldots, m_{n-1}$. As none of the m_i can be negative and their sum could not exceed M, the sample space W will be composed of points E with all coordinates $m_1, m_2, \ldots, m_{n-1}$ being non-negative integers and satisfying the inequality

$$m_1 + m_2 + \ldots + m_{n-1} \leq M \tag{58}$$

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

$$\Sigma (m_i - Mp_i)^2/Mp_i = \chi_\varepsilon^2 \tag{59}$$

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

$$x_1 = x_2 = \ldots = x_n \tag{60}$$

Having disposed of this we may go on and discuss the probabilities of errors. First of all: is it legitimate to discuss the probabilities of errors in testing statistical hypotheses? Isn't it equivalent to discussing the probabilities of hypotheses themselves, which would be useless? E.g., considering the Student's hypothesis, it would be useless to discuss its probability because this would be also the probability of $\mu = 0$. As μ is an unknown constant, the probability of its 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 is impossible to decide what is the value of $P\{\mu = 0\}$.

To this criticism that may be suggested 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_o, depends on the position of the sample point E, that is to say, on its coordinates, and those, 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 its 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 $\frac{1}{2}$. On the other hand if $\mu = 10$, say, the same probability may be equal to 0.0001. Therefore we shall agree to denote by $P\{E \in w|H\}$ the probability of E falling within w calculated on the assumption that the hypothesis H is true.

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

$$P\{E \in w|H\}$$

of E falling within w.

Now let us turn to errors of the second kind, defined on page 45. For an error of the second kind to be committed it is necessary (and sufficient) that the hypothesis tested be wrong and that the sample point fail to fall within the critical region selected. Therefore, the probability of an error of the second kind is

$$1 - P\{E \in w|H\} \tag{61}$$

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

$$\beta(w|H) = P\{E \in w|H\} \tag{62}$$

$\beta(w|H')$ 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 having regard to the properties of the power (62). In fact the values of the power $\beta(w|H)$ for a fixed region w and for a changing hypothesis H (which in particular may be H_0, the one we desire to test) describe no more and no less than the properties of the test based on the critical region w. In fact, what is it that could be called "the properties of a test?" To

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

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

1. The probability of rejecting the hypothesis H that $\mu = 0$, is always greater when this hypothesis is wrong than in cases when it is right. 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 tested when it is in fact wrong.

Details of the theory of testing statistical hypotheses based on considerations of the probabilities of errors will be found in the serial "Statistical Research Memoirs" issued by the Department of Statistics, University College, London. (First volume issued in 1936). Various articles in that publication contain numerous references to similar works published elsewhere.

- - - - - - - - - - - - - - - - - -

Editors note: The reader may be puzzled by the sudden appearance of $\beta(w|H)$ in Eq.(62); it may seem that $P\{E \in w|H\}$ could as well be used as a symbol for the power of a test. The point is that $P\{E \in w|H\}$ was introduced on page 16 to denote the probability of E falling within w, as determined by H, this probability being considered a function of **w**, H being fixed. It was called the integral probability law of the x_i as determined by H. On the other hand, $\beta(w|H)$ is considered as a function of H, the region w being fixed. To go into more detail, one writes $P\{E \in w|H\}$ if the emphasis is on the probability of E falling within w for a given set of parameters of the elementary probability law (page 16); in such a case one might be interested in seeing how the probability $P\{E \in w|H\}$ varies with w. On the other hand, one writes $\beta(w|H)$ if one is interested in seeing how various possible values of the parameters of the elementary probability law affect the probability of E being an element of a fixed region w.

RANDOMIZED AND SYSTEMATIC ARRANGEMENTS OF FIELD EXPERIMENTS

A conference with Dr. Neyman at the Cosmos Club in Washington, 7th April 1937, 2 p.m., Mr. Frederick F. Stephan, Secretary and Editor of the Journal of the American Statistical Association, presiding.

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

Those of the present audience who are familiar with the material of my first two lectures must be aware that the answer to the question must be both empirical and subjective. The application of formulas of mathematical statistics to the results of agricultural trials presumes the existence of some mathematical model of these experiments, and the question under consideration reduces itself to whether the correspondence between this model and what happens in actual practice is sufficiently accurate. This question is exactly similar to the one whether the formulas of plane geometry could be applied to measure this or that area on the surface of the earth. Another similar problem, also mentioned in my second lecture (page 24), is whether 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 it involves trials in conditions of actual practice. The subjective character is unavoidable, because having the results of the trials and also the corresponding theoretical deductions from their mathematical model, we have to 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 to measure all of them separately. If the measure of the whole appears to be very different from the sum of the measures of the 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 may vary.

Having this in view I am going to give you a short account of the work recently done by Mr. C. Chandra Sekar* in the Department of Statistics providing 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 that are contained in my second lecture (pp.19-32): 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 to have the trials randomized. But the question whether the agreement is satisfactory or not will be left to you.

Now I must enter into the details and describe the experiments that I have in mind. I shall deal with the experiments of a very common type in which plots are rather narrow and long rectangles all arranged in one row. They are combined in 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. Adding to it some details as to how 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. You know that in this arrangement each of the objects is repeated in each of the blocks the same number of times, e.g. once, and that the order in which the objects occur within each block is determined by random sampling. If the number of the 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 the 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 to calculate the estimates of the true difference between the mean yields which any 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. We know also how to calculate an un-

---

* Just where Mr. Chandra Sekar's paper will be published has not yet been decided.

biased 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 got some of the worse ones) the further theory is not entirely clear.* **

It is probable, however, that the application of the t test gives the results very much in accordance with its theory: 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 wrong with relative frequencies in good accordance with the mathematical tables.

Many practical agriculturists find that the distribution over the field of the objects compared, if left to chance, is not always satisfactory. 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

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, which were meant for randomized block experiments. There is no doubt that from the point of view of theory such a procedure is wrong. The theory of randomized blocks assumes specifically that the blocks are randomized and its validity is easily shown to depend upon this assumption. But it is a question how large are the discrepancies between theory and practice arising from the disregard of this condition.

The above systematic arrangement is very popular in Poland. I have 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¤ giving a

---

* J. Neyman with cooperation of K. Iwaszkiewicz and S. Kolodziejczyk, "Statistical problems in agricultural experimentation," Supplement to the J. Royal Stat. Soc., 2, 107-180, 1935.

** B. L. Welch, "On the z test in randomized blocks and Latin squares," Biometrika 29, 21-52, 1937.

¤ J. Neyman, "The theoretical basis of different methods of testing cereals, Part II: The method of parabolic curves." Published by K. Buszcynski & Sons, Ltd., Warsaw, 1929. Price, about 50 cents.

statistical theory of the agricultural trials arranged systematically.

The general lines are as follows. It is assumed that the natural level of fertility along the field may be adequately represented by a parabola of some not very high order, say the 4th. If u denotes the coordinate of the center of any of the plots, starting from the left, so that

$$u = 1, 2, \ldots, N \qquad (1)$$

then the true yield of A, if it were tested on the u-th plot would be

$$A(u) = A + bu + cu^2 + du^3 + eu^4 \qquad (2)$$

where A is a term depending on the object* A (treatment or variety), and b, c, d, and e are unknown constant coefficients. A is here used 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

$$B(u) = B + bu + cu^2 + du^3 + eu^4 \qquad (3)$$

where B again depends on the object B but the other constants b, c, d, and e are the same as in Eq.(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, 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** theorem to get the estimates of the differences like

$$B - A, \; C - A, \text{ etc.}$$

and of their respective variances.

---

* Dr. Neyman uses the word objects here to cover whatever is being compared in the experiment. The objects might be treatments, or varieties, etc. Special features exist in a comparison of varieties, other features exist for fertilizers. In the present discussion, the analysis applies as well to a comparison of treatments as to a comparison of varieties, and the word object is used to refer to either. A more felicitous choice might have been found, but inquiries have not yet brought forth anything better. Editor.

** See also F. N. David and J. Neyman, "An extension of the Markoff theorem on the least squares" in the Statistical Research Memoirs, vol.II (in preparation; published by the Department of Statistics, University College, London W.C.1)

Ten Lectures and Discussions on

SCIENCE: Its history, Philosophy,

and Relation to Democracy

held at the Department of
Agriculture Graduate School
under the chairmanship of

M. L. Wilson
Under Secretary of Agriculture

- - - - - - -

Edited by W. Edwards Deming
Assistant Director of the Graduate School

- - - - - - -

Lecture VII

SCIENCE AND DISCOURSE

by

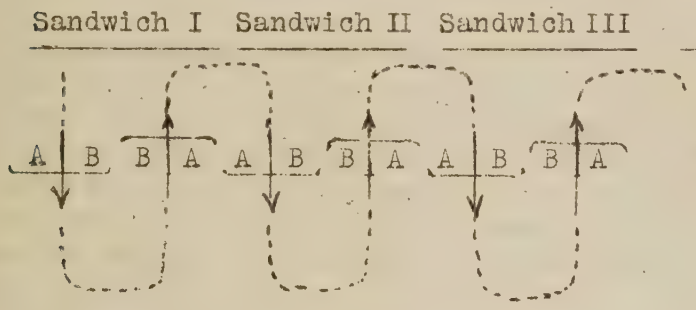
Charles W. Morris

Associate Professor of Philosophy
The University of Chicago

The Graduate School
The Department of Agriculture
Washington
1939

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

This was invented by Dr. E. S. Beaven* who used it with great success breeding his renowned varieties of barley. The half drill strip experiments are designed to compare only two objects A and B, say two varieties. They 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)

You see that 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 you will see that those blocks are not randomized.

We must distinguish here between two kinds 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 sandwich. 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 i-th 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 as described above. It is a matter of an easy adjustment of a few formulas

* E. S. Beaven, Jour. Ministry of Agriculture 29, 436-444, August 1922.

and of preparing tables to facilitate the calculations. But here again we come to the question whether the scheme underlying the method of parabolic curves corresponds sufficiently close to what happens in practice.

I shall now discuss the question of what empirical data are needed for deciding whether any particular mathematical model corresponds to the experiments.

When comparing any two objects A and B, of which the former 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_o that the true average yield $\overline{B}$ of B if sown on the whole field, does not exceed that of A, say $\overline{A}$. That is, H_o is the hypothesis

$$\overline{B} - \overline{A} \leq 0 \qquad (6)$$

Whichever of the mathematical schemes described is applied, the test of H_o consists (i) in calculating the estimate of $\Delta = \overline{B} - \overline{A}$, say x, (ii) in calculating the estimate s'^2/n of the variance of x, and (iii) in referring the ratio $t = \overline{x}/(s'/\sqrt{n})$ to Fisher's table of t. If the observed value of t exceeds the value of t_α corresponding to some small value of P, say 0.05 or 0.01, then the hypothesis H_o 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 the field trials must be randomized, whether the non-randomized trials give any sort of bias in the statistical tests, is reduced to the following:

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

(2) Whether, in cases when the hypothesis H_o is wrong and thus $\overline{B} - \overline{A} = \Delta > 0$, the t test detects this circumstance, the value of t falling above the critical t, with a frequency predicted by the theory, $\Delta = \overline{B} - \overline{A}$ having a prescribed magnitude.

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 $t = \overline{x}/(s'/\sqrt{n})$ does not correspond to the actual trials, and either the model or the experimental design should be altered. If, however, a considerable volume of

---

* At this point the reader may wish to refer back to pages 28 and 29.

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 when it is in fact true, and detects its falsehood when it is wrong, and with about the same frequencies as predicted by theory, so the experimenter knows where he is.

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

It is easier to obtain an empirical distribution of t for the case when $\overline{A} = \overline{B}$. 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 those 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. Not all of them, however, 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 would be the results of some particular experiment on this field, as in comparing some objects A, B, ..., 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 parallelogram represents a half drill strip

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

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. Those would be the actual 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 $\overline{A} = \overline{B}$. A number of such values of t, independently calculated, will produce the distribution we want to compare with the theoretical ones deduced by Student, namely

$$p(t) = C(1 + z^2)^{-\frac{1}{2}n} \qquad (7)$$

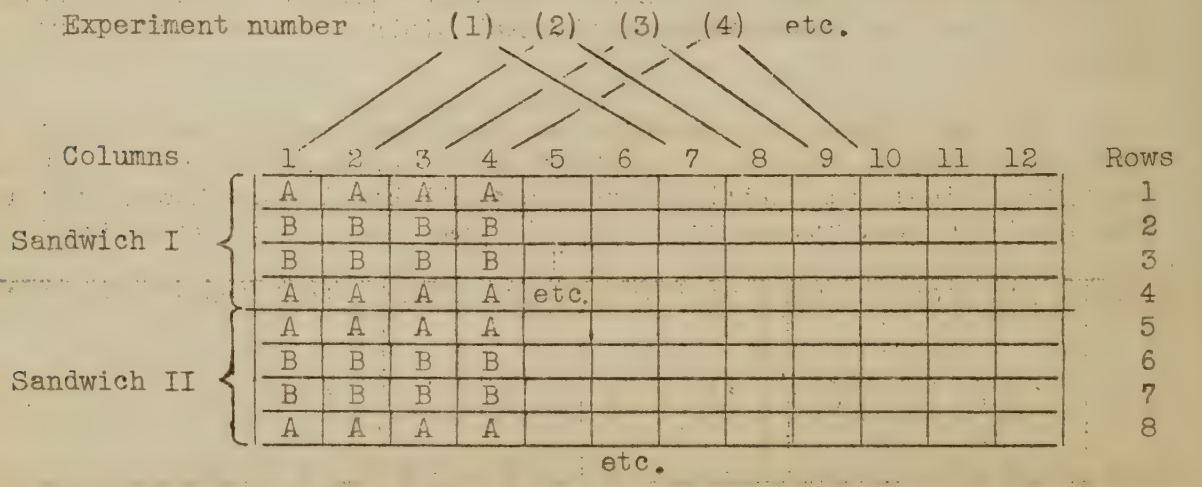
where $t^2 = z^2(n - 1)$, and $n - 1$ is the number of degrees of freedom on which the estimate s'^2 is based.

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

$$s'^2/n = \sum (x_i - \overline{x})^2/n(n - 1) \qquad (8)$$

The first authors to run tests on uniformity trial data to see whether the distribution of $\overline{x}/(s'/\sqrt{n})$ from non-randomized sandwiches follows Student's frequency of t, so far as I am aware, were S. Barbacki and R. A. Fisher.* They came to the conclusion that the lack of randomization is destructive to the t test, and they blame Student for thinking differently. It seems to me, however, that they were a little unfair to Student, and that the figures they produced are not sufficient to support their statements.

They took just one uniformity trial in which weights of yields of wheat on short parts of single rows were published.** They combined the adjoining rows to obtain the width of a half drill strip. The rows were sufficiently long and they divided them into 12 columns and so obtained 12 columns of hypothetical half drill strips, each being a contin-



* S. Barbacki and R. A. Fisher, Annals of Eugenics 7, 189-193, 1936.
** G. A. Wiebe, Variation and correlation in grain yields among 1500 wheat nursery plots. J. Agric. Res. 50, 331-357, 1935.

uation of the strips in other columns. Then they said that these columns will represent the results of six hypothetical experiments comparing some variety A with another B. Experiment No. 1 would consist of sandwiches in column 1 and column 7; experiment No. 2 would consist of sandwiches in columns 2 and 8, etc., as marked in the figure. Then they 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 of the t had the same sign! This of course should be expected, since the t 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 (see plan on page 56) 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 among themselves.*

It follows that we could not accept the results of Barbacki and Fisher as conclusive in the question we are interested in. Their figures emphasize only the known fact that there is danger in replicating an arrangement on plots in adjoining columns, as an error in one of them is likely to be repeated in the others. This does represent an advantage of the randomized arrangements but does not show that systematic experiments, if carried out with due precautions, give necessarily 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 his 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 the 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. There is no doubt again 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 the parabolic curves, is based exclusively on the assumption that the experiment is arranged systematically. Whether it works satisfactorily must be tested empirically.

- -

* Student, "On testing varieties of cereals," Biometrika 15, 271-293, 1923; see pp.286-287 in particular.

Some work designed to throw light on the question we are interested in 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 these experiments was 120. For each of them he calculated t, once using the formula implied by the new Student's method and next, applying the little more complicated procedure of the method of parabolic curves (pp.51-52). The two diagrams on the next page show the results.*

These are the objective results referring to question 1 above concerning the case where the compared objects are identical. They may be supplemented by the probabilities of getting something worse merely by chance. In the new Student's method this probability is 0.173 and with the method of parabolic curves it is 0.643. These figures seem to indicate a certain advantage of the method of parabolic curves.*

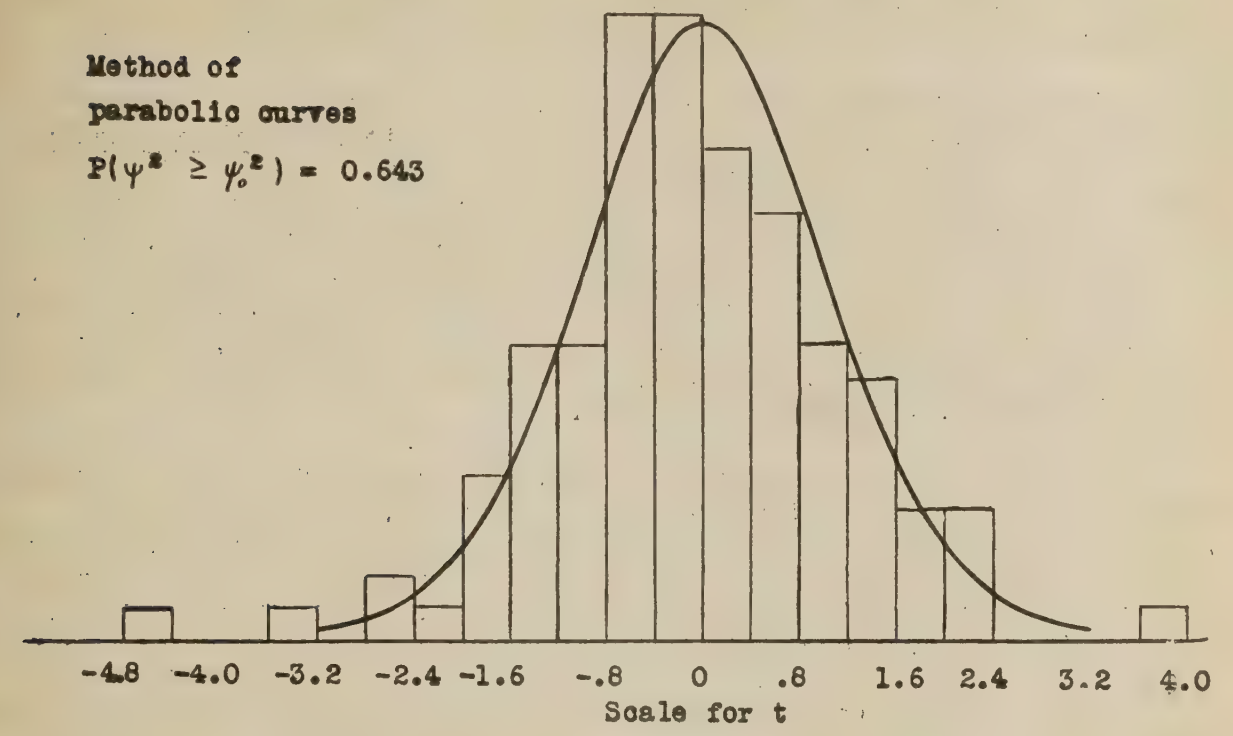
Having those objective empirical results, it is now a personal question whether to consider them favorable or unfavorable to Student's opinion that so far as the validity of the t test is concerned there is no special harm in the lack of randomization of the sandwiches. If you want my personal opinion on this point, it is this: were the lack of randomization of the sandwiches really disastrous for the t test in the case of the two compared objects being identical, then the available empirical material would have demonstrated this circumstance. Some divergence probably exists, and if more empirical data were available it would doubtless become apparent. However, the divergence could not be of very great importance. I always prefer to deal with mathematical models corresponding as closely as possible to practical experiments. Therefore, if the practical agriculturist insists on his sandwiches being systematically replicated, I would advise him to work them out using the method of parabolic curves rather than the new Student's method based on a hypothesis concerning the level of fertility, which seems a little artificial. The empirical results (the graphs on the next page) seem to show that this is safer.

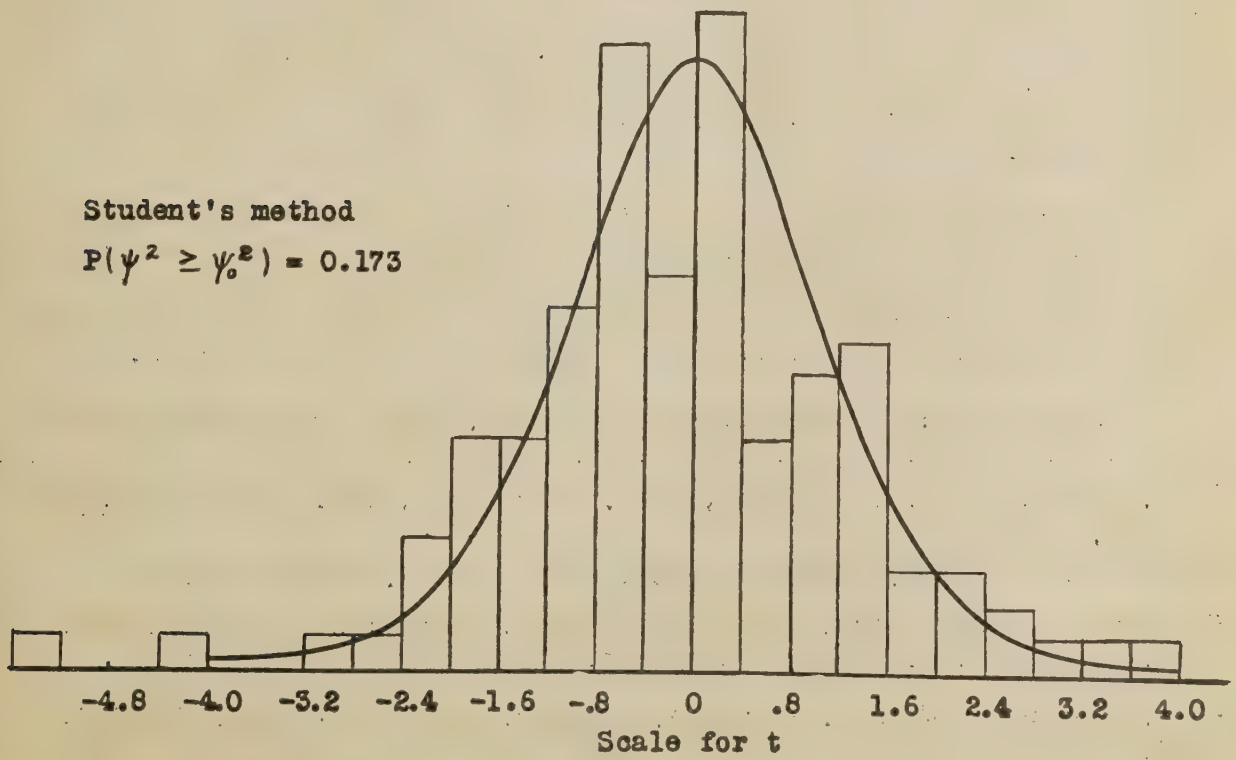
We may now turn to the next question and see the results of trials with systematically arranged sandwiches when the object B does give greater average yields than the object A. The theory of the t test referring to this case is not so generally known as the results arising from no difference between A and B, and I will remind you of certain important points.

---

* In regard to the quantity ψ^2 appearing in the diagrams on the next page, it may be said that this is a criterion for testing goodness of fit, which, for large samples, proves to be unbiased and most powerful with respect to the alternatives that are "smooth." Its theory was briefly described by J. Neyman in the Comptes rendus 203, 1047-1049, 1936 with a further note in the same volume on pp.1211-1213. The "smooth" test for goodness of fit will appear in the Skandinavisk Aktuariedskrift pp.149-199, 1937, (in English).

t DISTRIBUTIONS IN THE HALF DRILL STRIP EXPERIMENTS





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

(ii) The frequency of the t test failing to detect a difference $\bar{B} - \bar{A}$ when it actually exists and is equal to ρ times the true standard error σ of 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; see Lecture III, p.45. The first short table of this kind was published by S. Kolodziejczyk**. This was later supplemented in a joint paper by myself, K. Iwaszkiewicz, and S. Kolodziejczyk¤, wherein certain graphs are published, of which two are shown on page 61. Finally, a differently arranged table was published by Miss B. Tokarska and myself‡.

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 you work on. To make this diagram 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 per cent 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 thus $\rho = 20/8.16 = 2.45$. 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 per cent 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 those two 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 is to see whether the existing divergences from this model influence the validity of the theoretical conclusions.

With regard to point (i) raised above, there are unsurmountable

---

* J. Neyman and E. S. Pearson, Phil. Trans. Roy. Soc. of London A231, 289-337, 1933.

** S. Kolodziejczyk, Comptes rendus, Paris, t.197, 814-816, 1933.

¤ Neyman, Iwaszkiewicz, and Kolodziejczyk, Supplement to the J. Roy. Stat. Soc., Vol.II, 107-180, 1935; pp.133 and 134 in particular.

‡ J. Neyman and Miss B. Tokarska, J. Amer. Stat. Assoc. 31, 318-326, 1936.

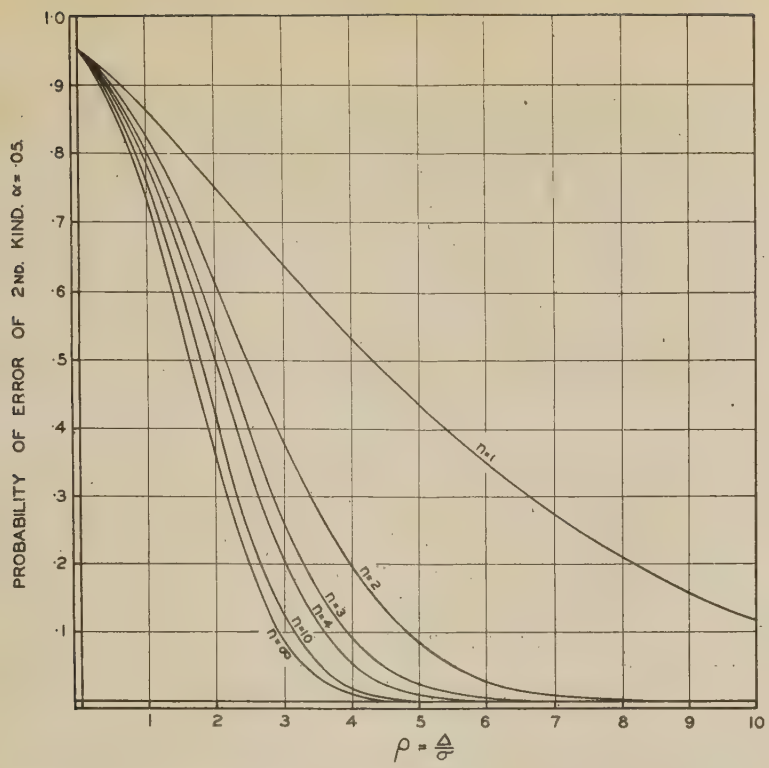


Diagram showing depenaence of probabilities of second kind errors on ρ *and* n, *when* $\alpha = 0{\cdot}05$.

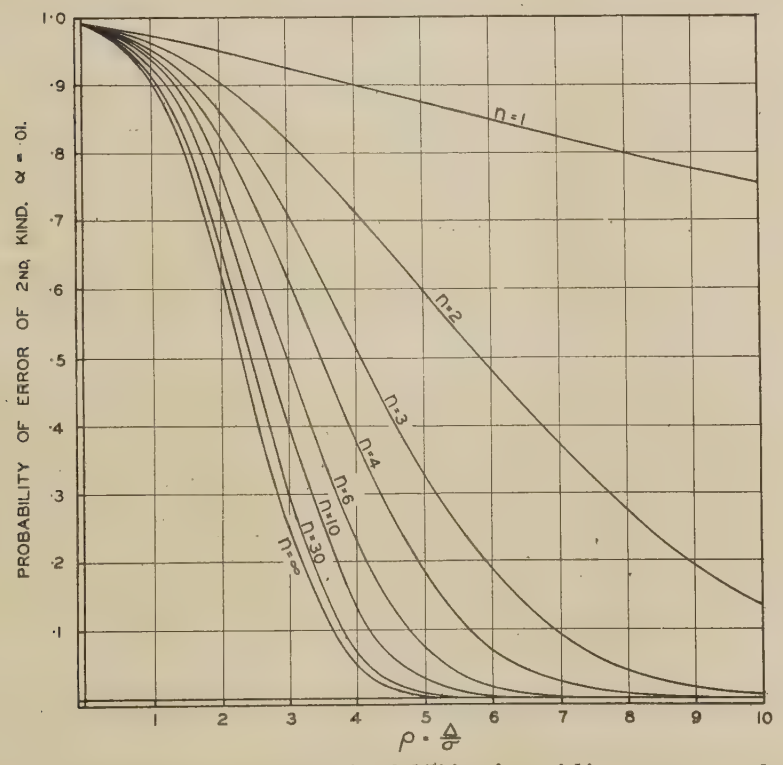


Diagram showing dependence of probabilities of second kina errors on ρ *and* n, *when* $\alpha = 0{\cdot}01$.

These diagrams are reproduced from pages 133 and 134 of an article "Statistical problems in agricultural experimentation" by J. Neyman, K. Iwaszkiewicz, and S. Kolodziejczyk, Suppl. Journ. Royal. Stat. Soc. vol.II, 107-180, 1935.

difficulties in this respect. There is no way to produce empirical evidence that in any fixed conditions of experimentation it is impossible to invent any test which 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 to t, could 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 empirically testing question (i).

The 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 by so much greater than that of A. For each such experiment we should 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 from the tables mentioned on page 60 or from the graphs on page 61.

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 55 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 clearer what was done next, consider the situation on any two particular fields. The assumed true standard deviations of the estimates of $\overline{B} - \overline{A}$ on those fields are respectively σ_1 and σ_2. Using the graphs of probabilities on page 61, 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. Those values were then 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

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

You will notice that $\Delta_1(20)$ represents the value such that if the difference between $\overline{B}$ and $\overline{A}$ 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 $\overline{B} - \overline{A} = \Delta_i(20)$.

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.

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

Theory	Method of parabolic curves	Student's method
percent	percent	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 whether the lack of randomization ruins the t test. The first column gives you 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 it 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 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 would be different. If we take the systematic arrangement of the blocks with more than two objects

ABCD, ABCD, ...,

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 little table gives a part of the uniformity trial data with rye as described by Hansen*. Looking at them you will hardly believe that all the plots were sown with the same variety and equally treated, but this is a fact.

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 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 <u>across</u> 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 system-

---

* N. A. Hansen, "Prøvedyrkning paa Forsøgsstationen ved Aarslev," Tidsskrift for Planteavl <u>21</u>, 553-617, 1914.

atic experiment replicated exactly in the two rows, e.g.

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 their direction is actually connected with that of ploughing. This, however, would require a considerable area.

In any case it seems advisable to carry out all cultivation processes common to all the plots, as ploughing, etc., in the direction across their greater length.

ON CERTAIN PROBLEMS OF PLANT BREEDING

A conference with Dr. Neyman in room 4090 of the Department of Agriculture, 7th April 1937, 10 a.m., Dr. S. C. Salmon presiding.

The problem that I am going to discuss refers particularly to the breeding of new varieties of sugar beets. However, it is probable that in the process of breeding other plants similar problems arise, and therefore the present discussion may have a wider interest.

The idea of the problem originated from contact with sugar beet breeders in Poland. The results that I am going to present, however, are due to Mrs. Y. Tang, M.Sc., all the details of which will soon be published in her paper prepared at the Department of Statistics, University College, London (cf. footnote p. 74).

The process of breeding new varieties of sugar beets is fairly complicated, but a rough idea of its essence could be obtained from the diagram on the next page representing schematically five distinct steps. In considering these steps we must remember several important points concerning the sugar beet. 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. Those are produced in the course of a second vegetative season when the plant uses the food accumulated previously 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 this makes it extremely difficult, if not impossible, to produce anything like a pure line. Finally we must remember that the production of new varieties may have various aims: we may try to produce beets with highest sugar content, giving the highest yield of roots per acre, or producing the highest yield of sugar per acre. The discussion which follows applies to the three cases, but we shall have in mind only the first of them.

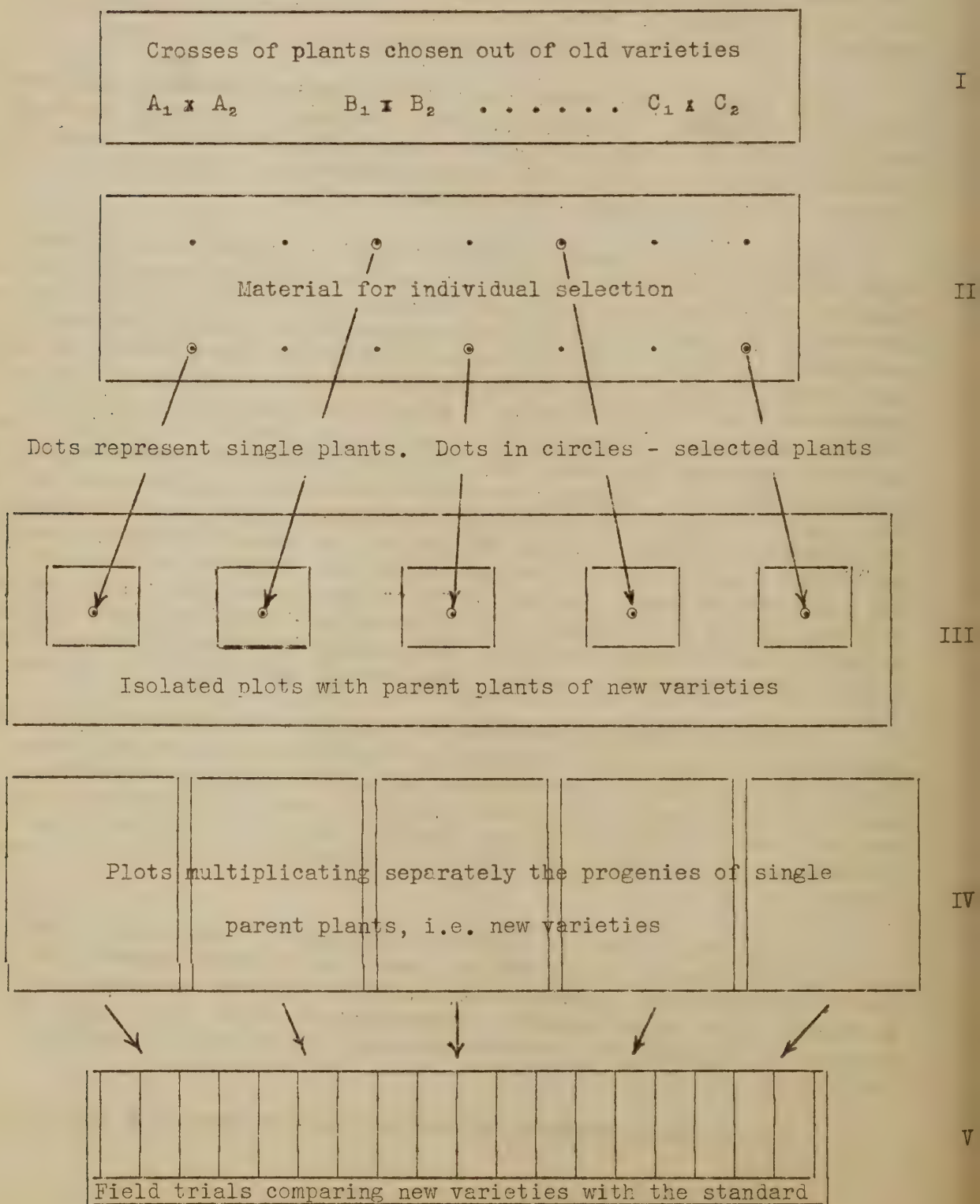
Having all this in mind let us consider the diagram and see what are essentially the five consecutive steps leading to new varieties. The first step consists in choosing out of the existing varieties a number of roots which, for various reasons, seem to be promising, and in forcing them to cross fertilize. For this purpose you plant those roots in pairs on plots, isolated from one another in a larger field of some cereal.

It is hoped that the capacity of producing high sugar content in old varieties may be cumulated as a result of crosses between them. But it is clear that a cross must sometimes cumulate 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

Plant Breeding

SCHEME OF PRODUCTION OF NEW VARIETIES OF SUGAR BEETS



and produce roots, forming the material for what is called the individual selection, the second step in our scheme. At the end of the vegetative period all the roots are lifted, washed, and weighed. Out of each root a small portion is cut out and analyzed for sugar content. This cutting does not kill the root, which is able to produce seeds as if it had been left intact. The majority of roots so analyzed are discarded as unsatisfactory. The remaining ones, with the highest sugar content, or having certain morphological characteristics indicating that they may be able to produce high sugar content, are stored for the winter, and then in the spring are planted separately on isolated plots to produce seeds. 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 checking whether and which of the newly bred varieties do possess any marked advantages 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 frequently to a greater extent, also on various conditions of environment.

Consequently 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 a small sugar content and unfortunately 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 also otherwise: analyses in V are a comparison of varieties, wherein the properties of individual roots are more or less ignored.

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

Anyhow, after the fifth step is concluded the breeder has to decide which of the new varieties are suitable for being put on the market. Other families 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 heterogeneous with respect to a number of pairs of genes. The consequence is that what we call a "new variety"

does not represent anything stable, but changes from generation to generation.

Further, according to the 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, without selection, from year to year. After a relatively short period the sugar content of new generations will drop low and he will lose the market. Consequently each breeder has to repeat constantly the steps described above - perhaps with certain modifications, and to start each year with step I, while continuing the following steps applied to varieties started 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. Those experiments are carried out in a number of places all over the beet growing districts of Poland. and all according to a certain fixed method, with the same number of replications, etc.

The seeds used are purchased on the market by a special committee and sent out to stations bearing conventional numbers but not the names of the producers.

The results are then officially published, and each of the producers can see how the results of his own efforts compare with those of the others. The consumers again are able to judge whether they were lucky in selecting the particular growers from which they bought the seeds. They try also to make some forecasts for the future. In this, however, they are frequently misled, because what is called a variety A produced by a firm X and put on the market in 1937 is essentially different from what will bear the same labels in 1938. Frequently it will also compare in different ways with another variety. Still, there is a certain amount of relative stability and the publication is useful.

DR. DEMING: Could you give some figures and the references which they and others like them are published?

DR. NEYMAN: The results of Polish competitive experiments are published yearly in Gazeta Cukrownicza, the official organ of the beet sugar industry. The following table gives the average results of the experiments of 1936.

DELEGACJA NASIENNA

Polskiego Przemysłu Cukrowniczego

Wyniki doświadczeń, wykonanych w roku 1936 nad produkcją buraków cukrowych z różnych odmian nasion

Przeciętne wyniki z 13 pól doswiadczalnych.

Doświadczenia wykonano w 120 powtórzeniach. Wyniki doświadczeń uwzględniono z mnożnikiem maksymalnym 5, zmniejszanym odpowiednio za punkty karne).

Nr. woreczka	Hodowla i nazwa nasion	Plon buraków z *ha* w *q*	kolejność	Plon liści z *ha* w *q*	Średnia waga buraka w *g*	% cukru w burakach	kolejność	Plon cukru z *ha* w *q*	kolejność
1	**Original Kl.Wanzleben - odm. ZZ**	306,63		119,0	354,6	19,15		58,72	
2	**Sand.Wielk.Hodowla Nasion**	*295,31*		*125,8*	*348,3*	*19,37*		*57,20*	
3	**Al. Janasz i S-wie - odm.AJ$_1$**	299,80		116,9	353,0	19,68		59,00	
4	**Original Kl.Wanzleben - odm.N**	366,81		125,6	422,2	17,75		65,11	
5	**Sp.Akc. Motycz - odm.MI**	307,32		125,6	356,2	19,37		59,53	
6	**Sp.Akc. Udycz**	307,30		133,3	356,9	19,59		60,20	
7	**Sand.Wielk.Hod.Nas. /powtórnie-kontrolne/**	*299,63*		*125,3*	*351,6*	*19,47*		*58,34*	
8	**Original Kl.Wanz. - odm.N. Nasiona zeszłoroczne, kontr.łańcuszkowa**	383,84		127.1	452,1	17,37		66.67	
9	**K.Buszczynski i S-wie - odm.NM**	314,91		117,8	369,0	18,94		59,64	
10	**Kaliska Hod.Nasion "Garbów"**	326,92		141,3	378,0	19,61		64,11	
11	**Sand.Wielk.Hod.Nas. /po raz trzeci - kontrolne/**	*295,98*		*124,4*	*348,3*	*19,33*		*57,21*	
12	**Original Kl.Wanzleben - odm.Z**	306,74		111,8	356,2	18,68		57,30	

The headings translated into English run like this.

Seed Testing Commission of the Polish Beet Sugar Industry

Results of the competitive trials with sugar beets of various origins ., carried out in 1936

Average results of 13 trials with the total number of 120 replicates

Bag No.	Breeder and variety	Yield of roots in quintals per hectare	Yield of leaves (same units)	Average weight of root in kilograms	Sugar content, percent	Yield of sugar in quintals per hectare

No. 8 translates into "variety N. Last year's seeds, chain control."

odm. = variety
1 quintal = 100 kilograms
1 are = 100 square metres = 0.02471... acres
1 hectare = 100 ares

The experimenters are not informed of the varieties of seeds contained in the bags.* Nos. 2, 7, and 11 contained the same variety, serving as a control of the accuracy of the experiment.

Similar publications exist also in other countries, but I can not give exact references. If I remember aright, in England the competitive

- -

* Dr. Neyman wrote to the editor as follows: "You may be surprised that the experimental results are printed, whereas the names of the varieties and the breeders are typed. This is because the latter information is kept in secrecy till the meeting where the accuracy of the experiments, and the agreement between the results of the experiments carried out in various places, are discussed. If the experiments disagree, then the commission may decide that the trials failed to provide reliable data, and in such cases they do not open the sealed envelopes containing both the order numbers and the names of the varieties; and the results of the trials are not made public." Editor.

experiments are carried out by the National Institute of Agricultural Botany in Cambridge. German results are, I think, published in the "Zuckerrubenbau."

As I have mentioned, the standard used by Polish breeders is always the variety that in the last year's competitive trials proved to be the sweetest. It changes from year to year.

It should be noted that this does not apply to the process of breeding of other plants. For example in Ireland, where most of the barley grown is bought by the Guinness Brewery, there is a perfectly established variety (barley is self fertilizing), namely the one that is preferred by Guinness. The aims of the breeders consist in surpassing this variety in certain of its properties, and it is used as standard in most (if not all) of the field trials. Later on I shall mention also some other important differences between breeding barley in Ireland and sugar beets in Poland, which have a direct relation to the problems that I shall discuss.

After this somewhat lengthy preliminary, we may turn to those problems. They are statistical in character and refer to steps II and V on page 68. 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. He may be unlucky in choosing plants for his crosses in I. But this is not a statistical problem.

2. Supposing he 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 p.68, and, consequently that some of his new varieties that actually come for comparison with the standard in V, are better than the standard. Obviously he may again be unlucky and lose those new varieties. The accuracy of the field trials is known to be limited and it is just possible that through unavoidable errors the experiments fail to detect the goodness of the best varieties and they will eventually be discarded. This, of course, would be most unfortunate, since it would mean a total waste of considerable amount of efforts, money, and time--a number of years! This 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 advices A and B are, of course, sound, but they will seem very troublesome to the practical breeder. His means are always more

or less limited and so is the area on which to carry out the field trials. Now each of the advices A and B, if followed, leads to an increase in the area of the field trials. And the breeder will ask: I am able to try out so many experimental plots on which to test a few new varieties, and in this case the trials will be with many replications and therefore fairly accurate; or with many new varieties, but then there will be only few replications and the test only superficial; which is better? Or rather: what is the right proportion between the number of new varieties to be started each year, and the number of replications in field trials comparing these varieties with the standard?

This is just the problem that was dealt with by Mrs. Y. Tang,* one of the students at University College in London. Her results show how to calculate approximately what would be 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 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 the Polish breeders goes into hundreds, while the number of replications they use is sometimes as small as four and I have not heard of its exceeding 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 perhaps 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 or in the other case misled.

I must now remind you of the nature of the errors that may be committed when testing statistical hypotheses. I will do so, treating 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 contents that those two varieties are able to yield if sown on the whole experimental field and if there were no technical errors. We are interested in the difference

$$\Delta = \overline{V} - \overline{S} \qquad (1)$$

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

---

* Mrs. Tang's paper is being printed in the forthcoming number of Biometrika, 29, Parts iii and iv, 1937.

experimental error ϵ, so that

$$x = \Delta + \epsilon \qquad (2)$$

Before putting the variety V 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, because otherwise his goods will frequently be inferior, and he will lose his customers. Mathematical statistics is helpful in this instance and provides means by which the frequency of cases when Δ is judged to be positive without being positive in actual fact can be reduced to any low level, α, called the level of significance and chosen in advance.

Statistically the problem of the breeder is reduced to the test of the hypothesis H_o that

$$\overline{V} - \overline{S} \equiv \Delta \leq 0 \qquad (3)$$

If as a result of this test we decide to reject the hypothesis H_o, 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_o will consist in the rule of rejecting H_o whenever

$$x/s > t_\alpha \qquad (4)$$

where s is the estimate of the standard error of x, and t_α is a constant number taken from Fisher's tables in accordance with the number of degrees of freedom on which the estimate s is based, and corresponding to his $P = 2\alpha$. 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 (and thus an unsatisfactory variety put on the market) only with the relative frequency equal to α. (ii) whenever H_o is true and the new variety is not so good as the standard, i.e. when $\Delta < 0$, this frequency will be even smaller than α. (iii) whenever the hypothesis tested 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.*

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 that will reach the market should not exceed 5 percent. Then he puts his level of significance at $\alpha = 0.05$, and finds in Fisher's Table IV the value of t corresponding to $P=2\alpha=0.1$. If

---

* J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses." Phil. Trans. Roy. Soc. London A231, 289-337, 1933.

the number of degrees of freedom is in his case 12, then t = 1.782. Thus he will reject the hypothesis H_o and say that his variety is good enough to be put on the market when $x > 1.782\ s$. Some other breeder B may consider 5 percent of all unsatisfactory varieties he gets too great a limit; he may consider that the proportion of such varieties slipping in on the market 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 his new variety through only if $x > 2.681\ s$. Some other breeder may be still more cautious.

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

DR. NEYMAN: Oh yes, there is, and I am most grateful for this question. The danger consists in the fact that whenever we are too particular in trying to avoid unjust rejections of the hypothesis tested, (that is, when it is in fact true), then we are exposing ourselves to an increased risk of failing to detect cases when V is actually better than S.

At this stage it will be convenient to use the special terminology introduced to distinguish between the two kinds of errors that we may make when testing a statistical hypothesis (Lecture III, p.45) 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 it is in fact 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, then he commits an error of the first kind. The error of the second kind consists in accepting the hypothesis tested when it is in fact false. Thus, when the breeder does not find sufficient reason for judging his variety satisfactory, i.e. when $x/s \leq t_\alpha$, when his new variety <u>is</u> actually sweeter than the standard, so that, though he does not know it, $\Delta > 0$, then he commits an error of the second kind.

The errors of the first kind are dangerous to the trade of the breeder, but so are the errors of the second kind. It must be remembered that each of them means a complete waste of efforts and money spent for a good number of years: after all those years a variety exceeding the standard in sugar content is successfully produced and then an error of the second kind causes it to be discarded. Therefore it is necessary to have as clear an idea as possible regarding the chance of committing an error of the second kind. For the numerical evaluation of such errors we use the charts on page 61, which were introduced for the study of the randomized and systematic experiments.

In the notation being used here, the "standardized" error of the second kind is

$$\rho = \Delta/\sigma = (\overline{V} - \overline{S})/\sigma \qquad (5)$$

This is the true value of Δ divided by the true value of σ, σ being the true S.D. of x (not the estimate s as used in Eq.4).

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, making n = 10 degrees of freedom. Suppose further that previous experience indicates that σ may be taken as something like 0.5. Let us now see what in these circumstances 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. To answer this question we calculate $\rho = \Delta/\sigma = 2$ and refer to the curves corresponding to n = 10 on page 61. It is seen that if we use the level of significance $\alpha = 0.05$ (the upper chart, page 61) then the probability of an error of the second kind is about 0.42. On the other hand, if $\alpha = 0.01$ (referring to the lower chart, page 61) 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 the new variety over the standard will be detected only in about 58 or 35 cases respectively out of a hundred. You see here 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 (65 out of 100 in place of 42 out of 100) of committing errors of the second kind.

It is obvious that the graphs on page 61 describing the dependence of the probability of errors of the second kind on the values of ρ and n are relevant from the point of view of problems in plant breeding considered here. In fact any seed breeding establishment, after a few years of its existence, must be aware of the size of the standard error per plot, say σ_o, which is likely to hold in future experiments. It is impossible to predict its exact value, but it is certainly possible to make rough estimates of its upper limit. Therefore the breeder contemplating experiments with m replications is able to substitute some reasonable number for σ into the expression for $\rho = \Delta/\sigma$, taking

$$\sigma = \sigma_o \sqrt{(2/m)} \qquad (6)$$

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

But in the case of the plant breeder a special difficulty arises. Suppose he finds that with 5 replications and $\alpha = 0.05$, the probability of detecting a good variety for which $\overline{V}$ exceeds the standard $\overline{S}$ by 5 percent is a fairly large one, 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 - IV (page 68) will yield him new varieties exceeding the standard in sugar content by so much as five percent. It is possible that such success in breeding is unthinkable, and that usually Δ does not exceed, say one-third of a percent.

Looking at the above 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, then practically all of his efforts in breeding new varieties will be wasted.

It is seen that the solution of the breeder's problem requires not only knowledge of the probabilities of errors of the second kind but also of the distribution that he is likely to obtain in the future for the values of Δ in the population of his new varieties. It is impossible to predict what will happen in the future, but it is possible to make rough guesses by studying what happened in similar circumstances in the past. We may try to estimate what was the distribution of Δ in past years and use those 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 the estimates of sugar excesses $x_1, x_2, \ldots, x_N$, and M estimates of corresponding standard errors $s_1, s_2, \ldots, s_M$. It is required to use those numbers to estimate the distribution, say $p(\Delta)$, of the true excesses $\Delta_1, \ldots, \Delta_N$ of those new varieties.

A similar problem was previously considered by Eddington and the solution is quoted by Levy and Roth,* but Mrs. Tang offers a new approach.** Her method consists in the following.

Denote by ν_k and μ_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 it be assumed that ϵ is normally distributed, which is a traditional assumption, then, as Mrs. Tang has calculated,

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

---

* H. Levy and L. Roth, Elements of Probability (Oxford University Press, 1936).

** The editor wonders if Mrs. Tang's method is not related to a scheme devised by Dr. W. A. Shewhart, "Correction of data for errors of measurement," Bell System Technical Journal 5, 11-26, 1926.

Mrs. Tang uses the assumption that σ has the same value in all the M experiments. This is partly justified by the fact that all of them are carried out on the same large field, by the same staff and with varieties having many similar properties. The common value of σ can then be estimated with great accuracy, being 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 v_1, v_2, v_3, v_4. Together with s' they will yield the estimates of μ_1, μ_2, μ_3, μ_4. Finally, having obtained the μ, Mrs. Tang uses them in a Pearson curve, which is then considered as an estimate of $p(\Delta)$.

It is difficult to test the efficiency of this method theoretically. But Mrs. Tang tried an empirical test. She started with an arbitrarily selected distribution represented by the histogram in Fig.1 on the next page. She considered the histograms as the true distributions of N values of Δ in some possible two experiments. Next she used Tippett's numbers or Mahalanobis' table* to produce normal deviates of x, such as might have been produced by N experiments satisfying her assumptions. In a similar way she obtained M values of the estimate of the error variance, each corresponding to one hypothetical experiment. Having obtained those quasi empirical figures, she applied her method to estimate the distributions of Δ. Fig. 1 shows the results obtained. It is seen that the continuous curves do agree with the "true distributions" represented by the histograms.

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

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

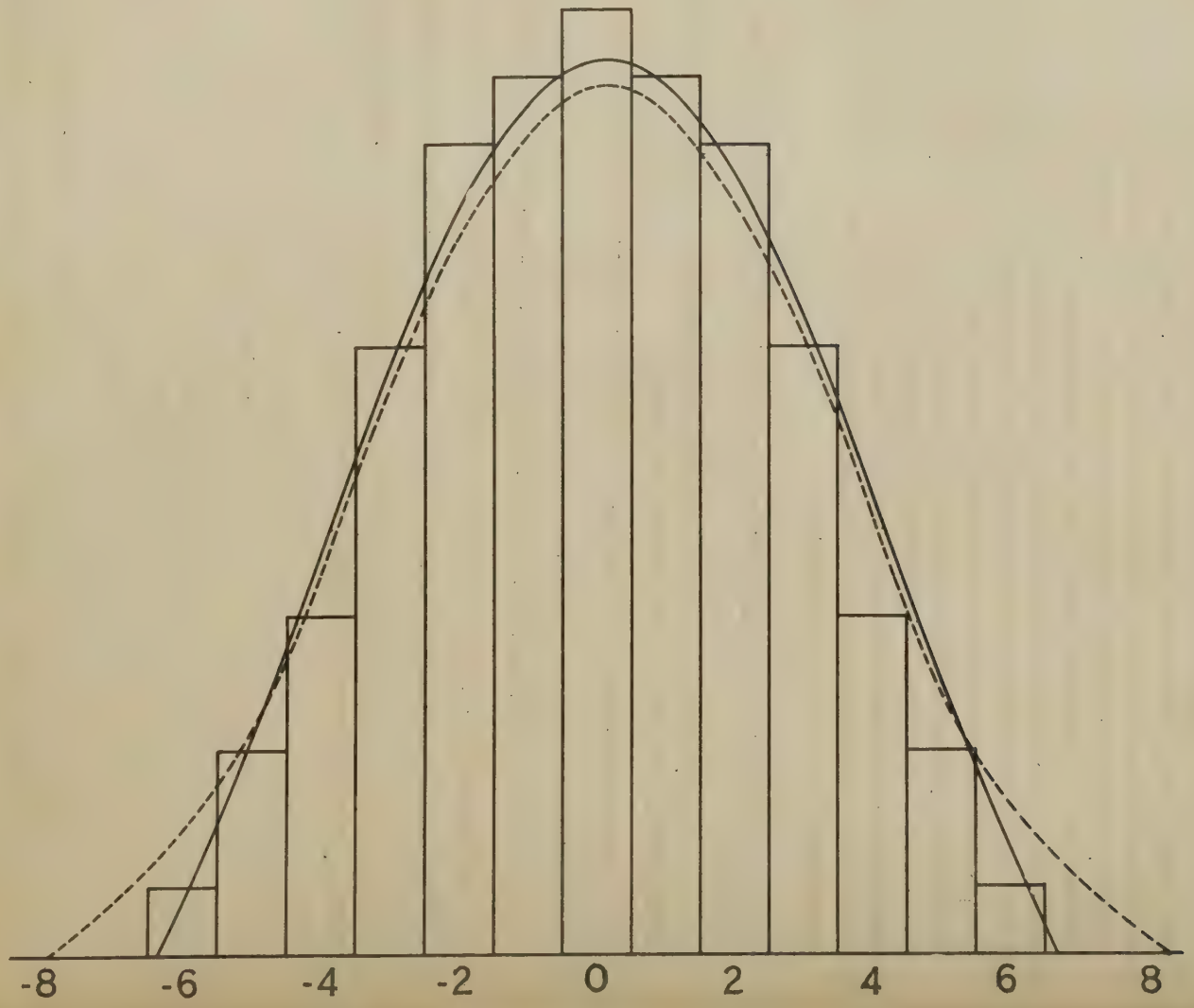
Value of Δ	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Frequency	1	3	5	9	12	13	14	13	12	9	5	3	1

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

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

---

* P. C. Mahalanobis, Sankhyā, The Indian Journal of Statistics (Calcutta) 1, 303-328, 1934.


HISTOGRAM
TRUE DISTRIBUTION OF Δ
ESTIMATED DISTRIBUTION OF Δ
ESTIMATED DISTRIBUTION OF x
-8
-6
-4
-2
0
2
4
6
8

Next take from the table of Mahalanobis (footnote page 79) the corresponding number of values of ϵ; he tabled them so that they may be considered as values of a normal variate about zero with unit standard deviation. Suppose that you find

$$\left.\begin{aligned} \epsilon_1 &= 0.03 \\ \epsilon_2 &= -1.16 \\ \epsilon_3 &= -0.25 \\ \epsilon_4 &= 0.53 \end{aligned}\right\} \quad (9)$$

etc.

Now add those numbers to your Δ_i and you will obtain what 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,

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

etc.

may now be used to estimate the distribution of x by the method of Mrs. Tang. Fig. 1 represents the results (page 80).

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, when 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 is still applicable with σ varying from one experiment to another within the limits that are likely to be met in practice; see Fig. 2 on page 83 showing the results with varying σ.

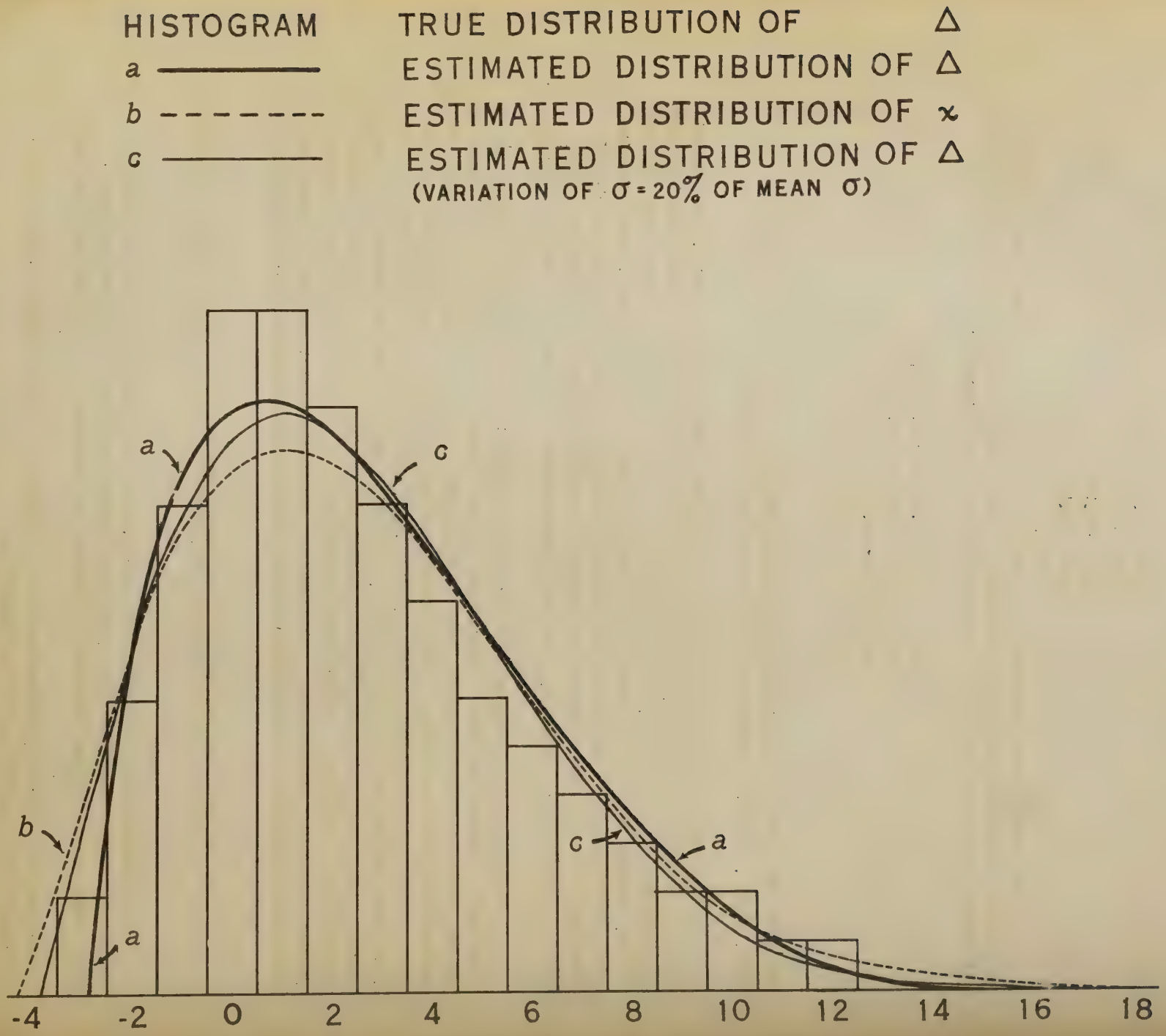
Having thus obtained an indication that her method does lead to reasonable results, Mrs. Tang applied it backward to estimate the distribution of true sugar excess over the standard in a number of new varieties tested in 1923 and 1924. The varieties were produced and tested by the breeders K. Buszczynski & 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. Those 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:

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

To work those experiments out, 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.* Next she estimated

---

* See conference on randomized and systematic experiments, pp.51-52; reference in footnote, bottom page 51.


HISTOGRAM
a
b
c
TRUE DISTRIBUTION OF Δ
ESTIMATED DISTRIBUTION OF Δ
ESTIMATED DISTRIBUTION OF x
ESTIMATED DISTRIBUTION OF Δ
(VARIATION OF σ = 20% OF MEAN σ)
a
c
b
c
a
a
-4
-2
0
2
4
6
8
10
12
14
16
18

the distribution of Δ, the true sugar excess. Fig.3 on page 84 gives the result referring to 1924. Here 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 various important information, which I shall classify under two headings.

1. He may use such curves to analyze his method of selecting parent plants in step II of page 68. Having records of how he selected them a few years ago, he may usefully study what would be the distribution of Δ if he had made his selection differently, say breeding only half of the families that he actually took. This would have allowed him to make a stricter selection of his parent plants, taking only the very sweetest. Ignoring the new varieties bred from the parent plants that in such cases would have been discarded, and estimating the distribution of Δ for the remaining ones, the breeder would be able to see whether taking many parent plants and breeding many new varieties does represent a marked advantage.

2. Having the estimated distributions of Δ corresponding to his actual experiments, and also to the stricter method of selection at step II, the breeder will be able to use the probabilities of errors of the second kind to see what would be the final results of his efforts including step V of page 68. Let us illustrate this for the estimated distribution of Δ given in Fig.3 page 84.

The breeder is naturally interested in those varieties for which $\Delta > 0$, called conventionally "good" varieties. Their proportion is represented by the area of the curve to the right of the origins of Δ (as in the curves of Figs.1, 2, 3). The breeder will be interested to know what proportion of these "good" varieties is likely to be detected as such by his field trials arranged according to this or that plan.

Take any positive value of Δ within the range of the curve in Fig.3, calculate the corresponding value of $\rho = \Delta/\sigma$ and use one of the graphs on page 61 to 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 such by the proposed trial.

Calculate $P(\Delta')$ for a number of successive values Δ' of Δ. Next, take the estimated ordinate $p(\Delta')$ of the distribution of Δ in the population of your new varieties (as an example, the full line curve of Fig.3). This multiplied by $\delta\Delta$ is approximately equal to the proportion of your varieties with Δ falling between Δ' and $\Delta' + \delta\Delta$. Multiplying, you will get

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

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

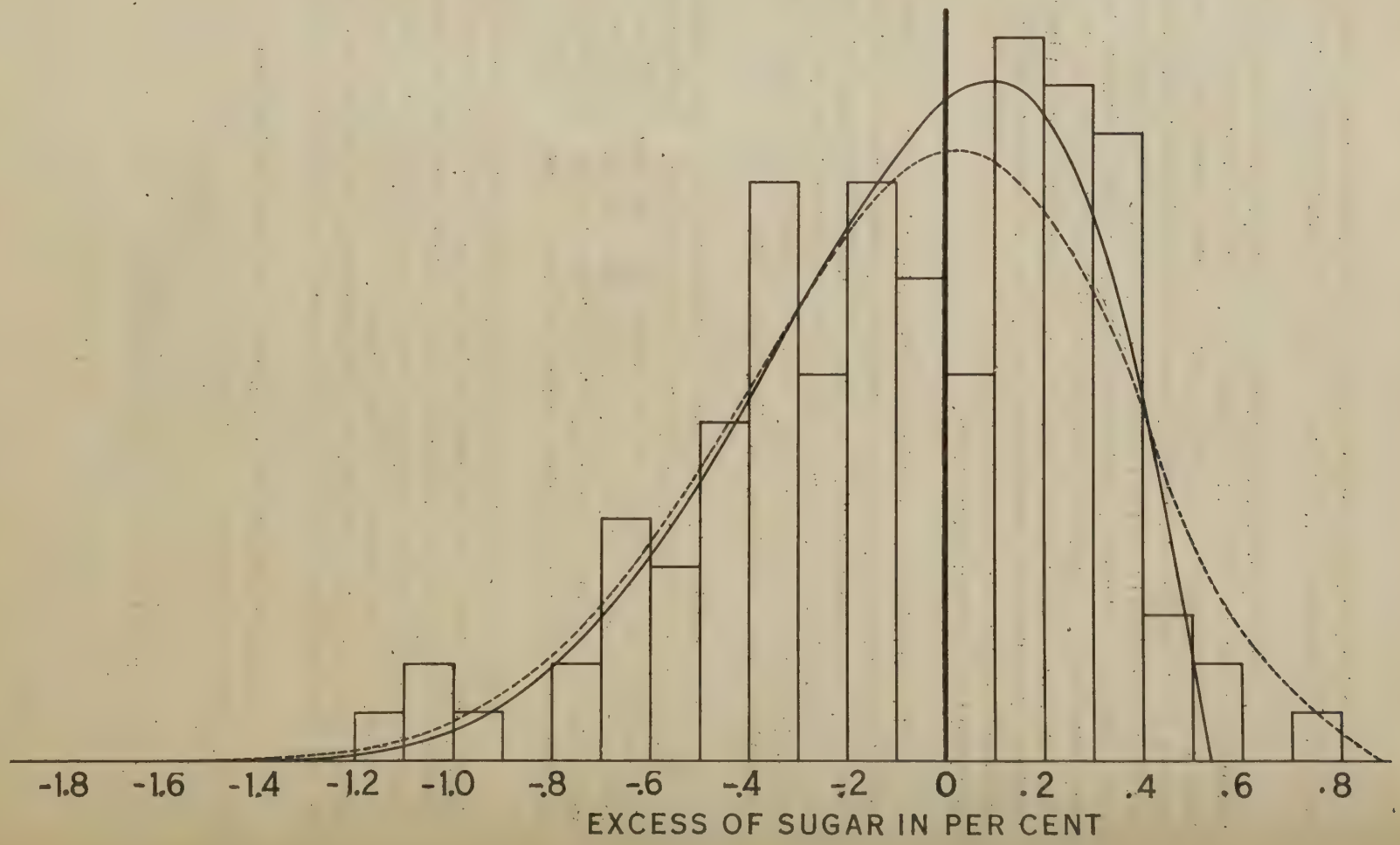


Fig. 3

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. Fig.4 was made up in this way from the "good" varieties of Fig.3 on page 84, i.e., it was made up from that part of the estimated distribution of Δ in Fig.3 lying to the right of the origin. The uppermost curve (a) of Fig.4 is simply the full line curve lying on the right of the origin in Fig.3. The dimensions are reduced so as to have the area under this part of the curve equal to unity: we are interested only in "good" varieties and in the proportion likely to be detected as such. In other words, the "good" varieties are here made the fundamental probability set of Lecture I.

All the lower curves (b and c) represent plotted values of the products $p(\Delta)\ P(\Delta)$, where $P(\Delta)$ corresponds to $\alpha = 0.01$ or 0.05, and to different arrangements of the proposed experiments. It was assumed that all these experiments would be arranged in randomized blocks and differ only in the number of replications m, marked on each curve. The curves corresponding to $\alpha = 0.05$ end at the axis of ordinates at the point 0.05. The other curves, 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 the given α and the given number of replications. Besides, 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 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 Fig.4, page 86. This number, 16.6 percent, is the area under the curve for $\alpha = 0.05$ 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 be termed conventionally the "best." There is no difficulty in calculating the proportions of the "best" varieties, the superiority of which 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 Fig.4. 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). The area

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 $\alpha = .01$

m = NUMBER OF REPLICATIONS

m	PROBABILITY OF DETECTING A GOOD VARIETY $\alpha = .05$	PROBABILITY OF DETECTING A GOOD VARIETY $\alpha = .01$	PROBABILITY OF DETECTING A BEST VARIETY $\alpha = .05$	PROBABILITY OF DETECTING A BEST VARIETY $\alpha = .01$
5	.343	.166	.542	.288
8	.438	.263	.696	.462
10	.485	.319	.767	.559
15	.558	.410	.865	.711
20	.609	.478	.922	.814

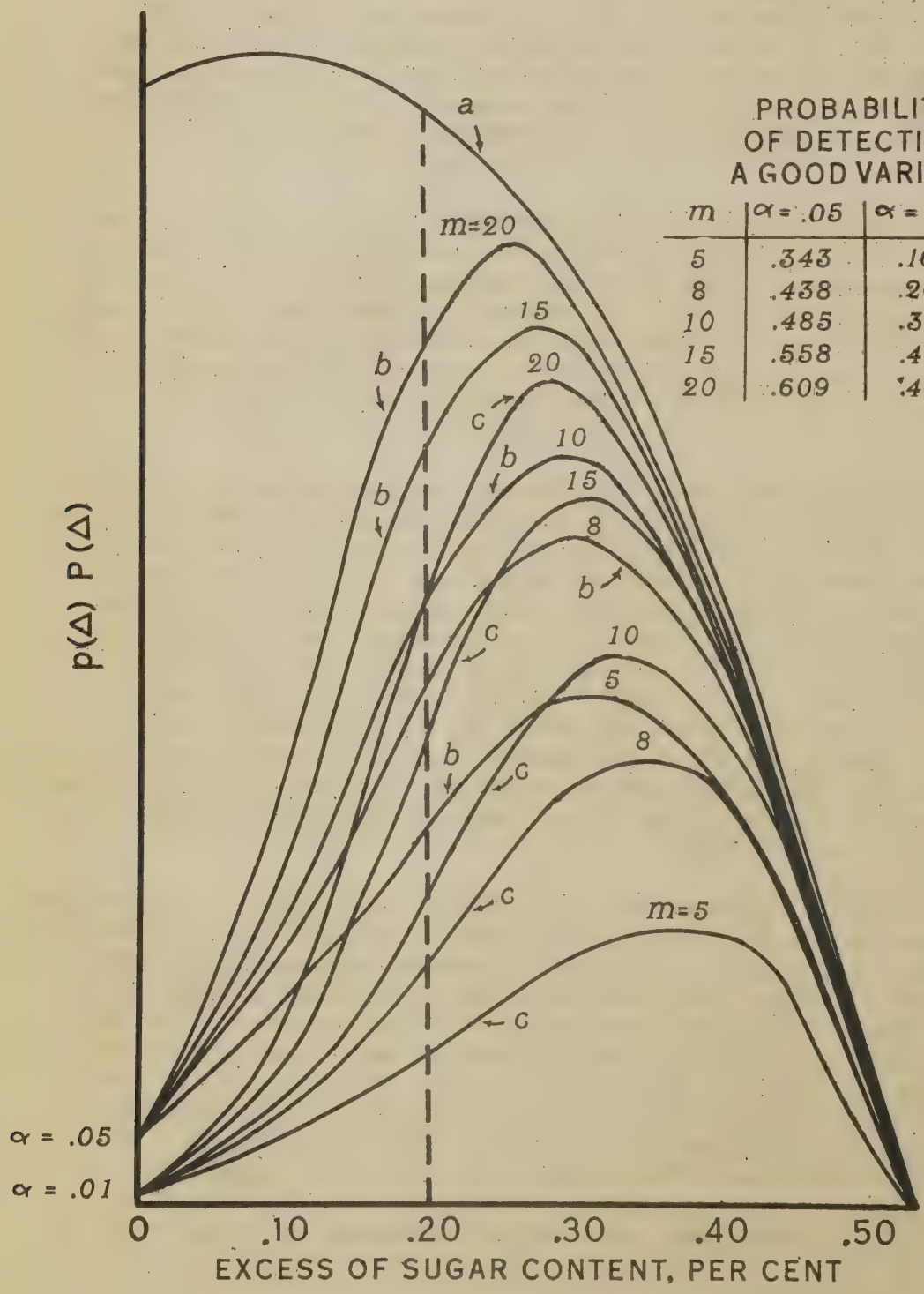


Fig. 4.

to the right of 0.2 percent under the curve(a)is now the fundamental probability set of Lecture I.

Fig. 4, the little table, and the method of their 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 Figs. 3 and 4 (pages 84 and 86), and compare the probable results of his work if the number of families started were not 500, but perhaps 400, 300, 200, with 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 presented above. In fact it is extremely difficult to include in formulas any process of more or less complicated practical work. Such is also the position in the present case. To give an idea of what I mean I may remind you of one thing I have already mentioned--new varieties are tested more than during one vegetative period and in more than one spot. It follows that the method as built up by Mrs. Tang refers to a simplified case. But it is obvious also that showing how to calculate the probable results of only one series of field trials, when no such method existed before, she does contribute something to our technique. And even if this is not all that is needed, it is really a lot, because the most difficult part of any problem consists in noticing that there is a problem at all and in advancing any sort of solution. There usually are a lot of people able to introduce the necessary corrections and extensions.

DR. SARLE: What basis do we have for figuring the possibility of including some false good varieties in this area (pointing to Fig. 4)? Will all poor ones be eliminated by this process, or is there a chance of getting some of the poor ones?

DR. NEYMAN: Fig. 4 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 of the best out of the "false good" varieties, those with $\Delta = 0$, to be passed as good, will be 0.05. On the other hand, the areas under sections of the curves in Fig. 4 give the proportions of those varieties that are really "good."

DR. SARLE: Your method does automatically that?

DR. NEYMAN: Yes, in principle; but we must remember that the method gives only an estimate, which is always liable to error.

DR. SARLE: How does it know which one to pick out?

DR. NEYMAN: It doesn't. It would be a great thing if it could. All it does 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 may safely say that, in the long run, the proportion of heads will be about $\frac{1}{2}$. 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."

DR. SALMON: This means with five replications you actually identify only a relatively small percentage of the total number of good varieties.

DR. NEYMAN: 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 indicated here.

---

* That is, if you toss it "fairly," which means to toss it so as to duplicate satisfactorily in a large number N of sets of n throws, the relative frequencies layed out by the binomial $(q + p)^n$, n being, for instance, 10, or any other convenient number, Cf. Lecture II, pp. 21-23 in particular. Just how this tossing is to be done is an experimental matter, but we have confidence that it can be done, because it has been accomplished in the past.

ON STATISTICAL METHODS IN SOCIAL AND ECONOMIC RESEARCH
Census by Sampling and Other Problems

A conference with Dr. Neyman in the auditorium of the Department of Agriculture, 8th April 1937, 8 p.m., Dr. Frank M. Weida, Professor of Statistics at the George Washington University, presiding.

I have received a number of questions for discussion at this meeting, and they lead me to believe that what I originally planned to talk about, namely some results of a particular research connected with systems of social health insurance, should be omitted. The questions asked are of extreme importance and I think they will take just the time that is available. I tried to classify those questions, and I shall try to answer them in groups. But it may be that such a collective answer will not be sufficient and then you will please simply ask additional questions.

There is a group of questions concerning the method of sampling; I think this is a very important question and I shall dwell some time on it. The typical question is how to get a good sample that will give sufficient data to estimate, say, the number of unemployed, the amount of money spent by the unemployed on certain kinds of commodities, and so on. There is a certain sum available for the inquiry and we have to decide what is the best way to use the money to obtain a good sample.

One particular question was asked referring to 300 cities, and the question was how to take a sample of them. It is suggested in this question that out of the 300 cities some 25 should be selected to represent the whole and that in each of the cities selected an exhaustive inquiry (complete census) should be carried out. The enumeration of all the workers in these 25 cities, of all the unemployed, the averages of moneys spent by them on various commodities, etc., should then be used to judge what happens in the 300 cities as a whole. I am expected to answer the question how best to select the sample of 25 cities that will be used for the above purpose.

I shall not answer this question. Instead I am going to advise as strongly as I can to drop the proposed method of sampling altogether. It is most dangerous and is practically certain to lead to deplorable results. By this, of course, I do not mean that a successful inquiry by a sample is impossible. On the contrary, my opinion is that the sampling method may be most useful and may provide very accurate results. What I emphatically protest against is the selection of any 25 cities for a complete census (as of the whole population, or of the employed or the unemployed, etc.), with a total omission of the remaining 275 cities.

Broadly speaking, there are two essentially different methods of sampling that are 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 is used in a special report* on the method, presented to the International Statistical Institute, it is generally accepted.

The method consisting in a selection of 25 cities out of 300 of them and in limiting the investigation to those 25 cities only falls under the heading of "purposive selection." The mere question addressed to me, how those cities should be best selected, suggests that the selection was not meant to be random, at least not entirely random. Usually it is suggested that the sample of the cities should be so selected that the averages of certain characters, called controls, calculated for the sample and for the universe should be in an as close agreement as possible, this circumstance justifies the term "purposive selection." But it is not the limitation of the randomness of sampling that makes the method dangerous. In fact, if it were only the question of 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" the total number of such units that might be selected must necessarily be small, and therefore the units themselves must be rather large. In your case you 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 essential difference between the two methods as 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) small groups (e.g. families, inhabitants of single houses, blocks, etc.) of which around 1000 or more are selected to form a sample, either entirely at random or at random with some restrictions.

The first method is hopeless, the other extremely useful. Those of you who would like to see theoretical reasons for this opinion, will find them in an article of mine.** Here I will give you an intuitive illustration of the ideas experienced there. Suppose we have a hundred

---

* L. A. Bowley: "On the precision attained in sampling." Bull. Int. Stat. Inst. 1926.

** J. Neyman: "On the two different aspects of the representative method," Journal of the Royal Stat. Soc. 97, 558-625, 1934.

dollars that we decide to use for gambling, with fair play. If we divide the whole sum into say five parts of $20 each and bet only five times, it is impossible to make any reliable prediction of what may be the result. We may lose all our money, or equally easy, we may double it. On the other hand, if we make a hundred bets at $1 each, then we may 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 around $20. The accuracy of the prediction would be still greater if instead of making a 100 bets at $1 we would bet a dime 1000 times.

Those 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 the total of 300 is based not only on theoretical considerations; some practical experience is available to show what might be the result of an inquiry if this method is applied.

In 1926 or 1927 two Italian statisticians, Gini and Galvani,* 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 it was necessary to clear the room for the new data; the old data were to be destroyed, but the statistical office considered it useful to have a representative sample from these data of the previous census, because it seemed just possible that in the future some new problem would arise, and it would be convenient to have the material. Therefore they decided to take a sample from the old data that could represent the situation in the whole of Italy.

They considered carefully the problem of how to obtain a better sample, and took into account the report to the International Statistical Institute. After a certain amount of discussion the Italian scientists decided to apply the method of purposive selection. The whole of Italy was divided into 214 administrative districts called circondarî. Those districts are large, some of them containing over a million inhabitants. Out of those they decided to select a sample of 29. You see, the size of their sample was larger compared with the universe, than in the case we were discussing, viz. 25 towns from a population of 300 towns.

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 controls into essential and secondary. They tried to select the 29 circondarî so as to have the

---

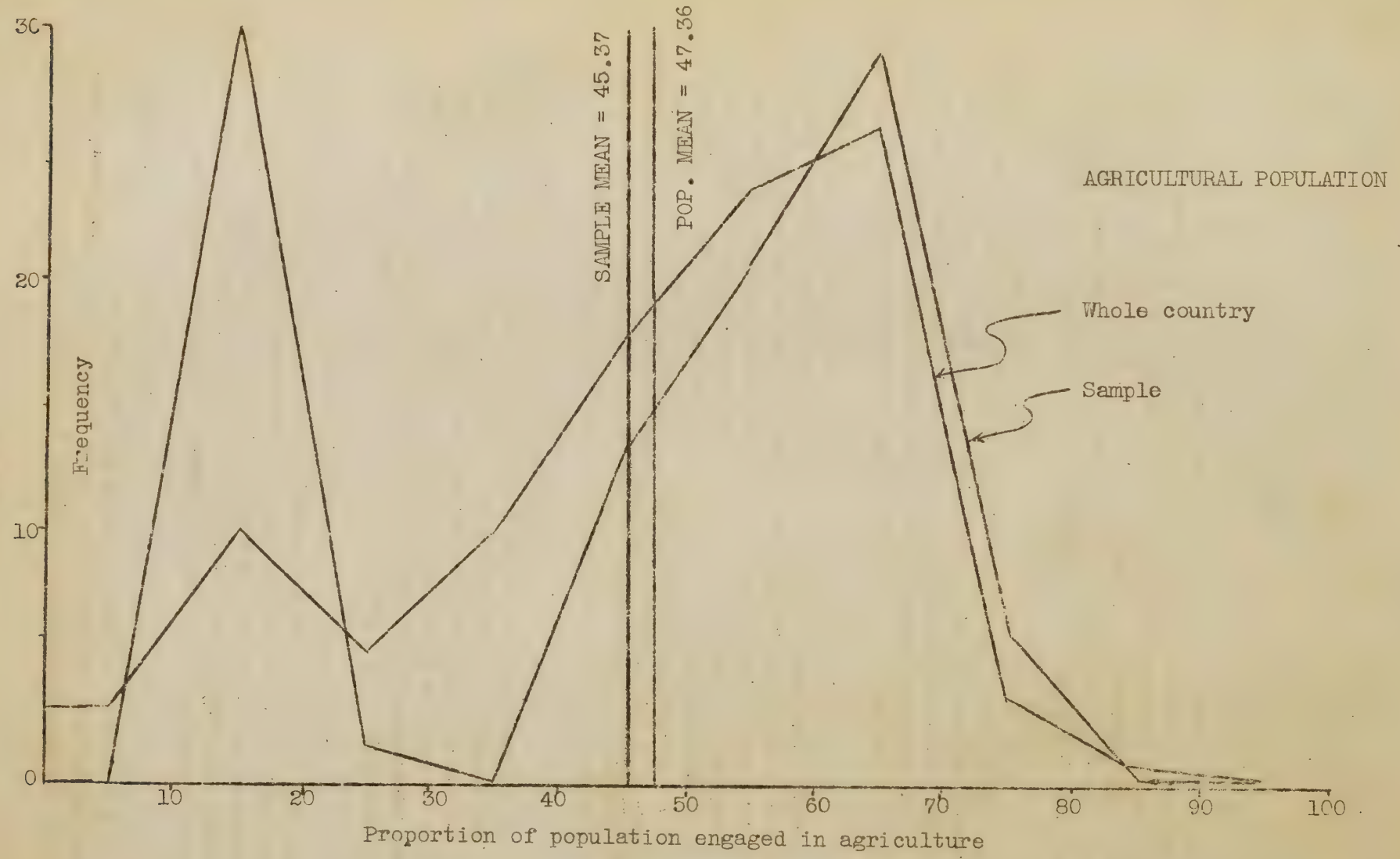
* Corrado Gini and Luigi Galvani: "Di una applicazione del metoda rappresentativo all'ultimo censimento italiano della popolazione," Annali di Statistica, Serie vi, vol.4 (107 pages), 1929.

means of the essential controls calculated from the sample practically identical with those for the whole population. They tried also to reach a reasonable agreement between the population and the sample means of the secondary controls. If you look at the figures, you will find that the agreement between the means of all the controls in the sample and the means of the same controls in the population is very good. I don't know exactly what happened next. I have the impression that the statistical office discarded the rest of the material and kept the sample. However that may be, the authors tried to see whether the sample they had selected was a success and whether it showed satisfactory agreement with the population also in other respects, besides in the averages of the controls. The result was very bad. They found that the distributions and also the correlations, in fact all characters except the means of the controls, as found in the sample, were in extreme disagreement with those in the population. The diagram on the next page, one of the many diagrams that the Italian scientists published themselves, gives an idea of the disastrous results that are apt to follow the sampling of big units. The proposed method (page 89) of selecting 25 cities out of 300 is likely to produce a similar result.

Having discovered that their sample of 29 circondarî is not representative of the whole population at all, the Italian statisticians expressed the opinion that it is generally impossible to obtain a sample that, as it were, would reproduce the sampled population with all its properties. Strictly speaking, of course, they are correct. There was in Italy in 1926 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 could 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 any really statistical problem of estimating means of any size, regressions, etc., a sample properly drawn, is for all practical purposes sufficient.

Now let us consider what is to be done to get a reliable sample. We must here 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 samplings, or great numbers of units that we draw separately. The sample of 25 cities or that 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.

It follows that instead of dividing your population into 300 parts, each inhabiting a particular city, you have to carry the subdivision much further. Probably the best thing would be to divide the whole into small groups inhabiting single houses or blocks. All those groups, which I shall call units of sampling, or simply units, must be listed, and this necessity of listing usually provides a limit to the tendency of having the units as small as possible. When this is done,



(Taken from page 95 of Gini and Galvani's article in the Annali di Statistica, Serie vi, vol.4, 1929)

you will be able to select a random sample of the units, which may be one - twelfth of the whole, like in the contemplated sample of 25 cities out of 300, but probably could be much smaller, without any serious detriment to the accuracy.

The process of random sampling may be of various forms, which are not indifferent from the point of view of accuracy of the results. The first attempt at a serious study of the relation between the method of sampling and the accuracy of the results was made by Bowley and is described in his report to the International Statistical Institute already mentioned. The main results are as follows.

The 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 in the case of the population formed by the inhabitants of the 300 cities, an unrestricted sampling combined with bad luck can produce a sample composed only of elements from 25 cities with the complete omission of the others. This, however, is extremely unlikely.

More accurate results could be obtained by what Bowley calls stratified, and what I call stratified proportional sampling. This consists in a two-fold subdivision of the population studied. We first divide it into a conveniently great number of larger parts, called strata. Those may be your 300 cities or some 600 halves of the cities, etc. Next, each stratum is divided into units of sampling. If it is decided to work with the sample of one-twelfth, then you select at random one-twelfth of the units out of each stratum separately. This makes it impossible for the sample to be devoid of the units representative to larger sections of the population.

It is obvious--and this presumption is supported by theory--that the more homogeneous the single strata, the better the effect of stratification. Therefore, if a city is divided into two or more parts, one inhabited by the well-to-do, the other by poorer people, still another being a shopping district, etc., all those parts should be treated as separate strata.

It may be useful to emphasize here that homogeneity of a stratum does not necessarily mean equality or similarity of all the people inhabiting or forming a stratum. In fact homogeneity of a stratum or of a population means a comparative similarity of the units of sampling. If the population of a town is composed of representatives of 10 different races all in the same proportions, then probably we should say that from the racial point of view that population is very heterogeneous. However, from the point of view of sampling it will be ideally homogeneous if it happens that the racial composition of any of its units is exactly the same as that of the whole population. It is seen that the internal heterogeneity of sampling units goes together with an external homogeneity of those units within the population. This is a general rule.

It follows that the choice of the units of sampling of a fixed size is not indifferent from the point of view of the accuracy of an investigation by sample. Mr. Frederick F. Stephan tells me that an investigation has shown a greater similarity between the inhabitants of two sides of one street than between those of the opposite sides of the same block. It follows from what I said that if it were contemplated to divide the population into units of sampling alternatively 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 a greater accuracy of sampling.

The gain in accuracy due to stratification is considerable, but it is possible to go beyond what was advised by Bowley. 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 imaginable. You could not expect that all the strata will be internally equally homogeneous. To make the situation clear, suppose that one of the strata, A, is ideally homogeneous, while some other, B, is fairly heterogeneous. Then, to know all about the stratum A it will be sufficient to take out of it one unit of sampling only. 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 could be put into exact form* and the result is as follows.

Denote by $\bar{U}$ the average referring to the whole population that it is desired to estimate from the sample with the greatest possible accuracy. Suppose for example that $\bar{U}$ is the average income per family of the unemployed. Denote by s the total number of strata, by M_i the total number of units of sampling forming the ith stratum, by n_{ij} the number of unemployed families within the jth unit of sampling belonging to the ith stratum, and finally by u_{ij} the total of the incomes of those n_{ij} families. With this notation we shall have

$$\bar{U} = \sum_{i=1}^{s} \sum_{j=1}^{M_i} u_{ij} \Big/ \sum_{i=1}^{s} \sum_{j=1}^{M_i} n_{ij} = \sum_{i=1}^{s} \sum_{j=1}^{M_i} u_{ij} \Big/ N \qquad (1)$$

Introduce further

$$\sigma_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (u_{ij} - \bar{u}_i)^2 \qquad (2)$$

where

$$\bar{u}_i = (1/M_i) \sum_{j=1}^{M_i} u_{ij} \qquad (3)$$

---

* See my article in the Journal of the Royal Statistical Society already referred to on page 90.

and denote by m_i the number of sampling units that we decide to select at random from the M_i possible units in the ith stratum.

It is possible to show that the greatest accuracy in estimating the numerator in formula (1) is attained when the numbers m_i of elements actually drawn in the sample from the M_i possible sampling units in the ith stratum are proportional to the products $M_i\sigma_i$, that is, when

$$m_i = m_o \frac{M_i\sigma_i}{\sum M_i\sigma_i} \tag{4}$$

so that the ratio $m_i/M_i\sigma_i$ is constant for all the strata, i=1, 2, ..., s.

I have denoted above by $u_{i1}, u_{i2}, \ldots, u_{iM_i}$ the totals of incomes of the families of the unemployed belonging to particular sampling units within the ith stratum. Denote now by $x_{i1}, x_{i2}, \ldots, x_{im_i}$ those of the u_{ij} that correspond to units actually included in the sample of m_i units drawn from the ith stratum, and let $\bar{x}_i$ be their mean. If the denominator in (1) is known, which is frequently the case, then the average $\bar{U}$ will be estimated by

$$\bar{X} = (1/N) \sum_{i=1}^{s} M_i\bar{x}_i \tag{5}$$

where

$$N = \sum_{i=1}^{s} \sum_{j=1}^{M_i} n_{ij} = \begin{matrix}\text{the denominator} \\ \text{in Eq.(1)}\end{matrix} \tag{6}$$

is the total number of the families of the unemployed within the population studied. The squared standard error of $\bar{X}$ is given by

$$N\sigma^2 = \frac{M_o - m_o}{m_o} \sum_{i=1}^{s} M_i\sigma_i^2$$
$$+ \sum_{i=1}^{s} m_i \left[M_i\sigma_i/m_i - (1/m_o) \sum_{i=1}^{s} M_i\sigma_i\right]^2$$
$$- (M_o/m_o) \sum_{i=1}^{s} M_i \left[\sigma_i - (1/M_o) \sum_{i=1}^{s} M_i\sigma_i\right]^2 \tag{7}$$

where

$$M_o = \sum_{i=1}^{s} M_i \text{ and } m_o = \sum_{i=1}^{s} m_i \tag{8}$$

are the total numbers of sampling units forming the whole population and the whole sample respectively.

It will be seen that of the three terms in (7), only one depends

on the numbers m_i of the units of sampling selected from particular strata. Therefore, any change in these numbers may influence only this term, which has the minimum value of zero whenever the m_i satisfy the condition (4).

If the denominator N in Eq.(1) is not known, then $\overline{U}$ will be estimated by a ratio of estimates of the numerator and the denominator separately. Some of the questions addressed to me refer to this situation. The estimate of accuracy of the estimate of $\overline{U}$ now involves the applications of some remarkable theorems of S. Bernstein* and of R. C. Geary,** but these matters are a little too complicated to describe here, and I shall have to refer those interested to my article in the Journal of the Royal Statistical Society already mentioned, page 90.

The adjustment of the numbers of samplings to be carried out within a single stratum is particularly important whenever we know from some a priori grounds that certain strata are more heterogeneous than some others. Such will be the case if some of your strata are cities with very mixed populations not permitting partition into more homogeneous parts, while other strata are uniform agricultural districts. In the light of formula (7), the purpose of stratification becomes now a little different from what it was before. Previously, it consisted only in getting strata as homogeneous internally as possible, though differing between themselves. Now we have an additional means of increasing the accuracy of the results by isolating into separate strata such parts of the population as are heterogeneous and by sampling them more heavily than the others.

It is obvious that the adjustment of the m_i to conform to Eq.(4) requires knowledge of σ_i. Those are never known before the investigation, otherwise the investigation itself would be unnecessary. The difficulty may be overcome in many ways:

1. When, apart from intuitive feeling, we have absolutely no previous knowledge concerning the variability of particular strata, it is useful to divide the inquiry planned into two parts: (a) preliminary investigation, and (b) the main investigation. The preliminary investigation consists in drawing at random very small samples, of n = 20 or 30, out of each stratum and using them to estimate the σ_i. If the estimate of σ_i is denoted by s_i, then the total number m_o of samplings intended should be divided between the strata in proportion to $M_i s_i$ so that

$$m_i = m_o \frac{M_i s_i}{\sum M_i s_i} \qquad \text{(compare with Eq.4)} \qquad (4a)$$

The main investigation would consist in selecting an additional m_i-n units

---

* S. Bernstein, "Sur l'extension du théorème limite du calcul des probabilités," Mathematische Annalen 97, 1-59, 1926.

** R. C. Geary, "The frequency distribution of the quotient of two normal variates," J. Royal Stat. Soc. 93, 442-446, 1930.

out of each stratum. It will be remembered that the preliminary inquiry will be useful also from the point of view of training the enumerators. Let me emphasize that the sample of the preliminary inquiry need not be large. In this respect see P. V. Sukhatme: "Contribution to the theory of the representative method," J. Roy. Stat. Soc. Supplement Vol.II, 1935, pp. 253-268.

2. Frequently we may have some previous knowledge of the strata considered. For example, in the process of working out the data of some previous general census, certain characters of the same units of sampling may have been calculated. Alternatively, the data for such calculations may be available. Those might not concern the character U, but it will be sufficient if we have information concerning the variability of the sampling units with respect to some character, say v, correlated with U. An adjustment of the numbers of samples according to the variability of v will be more or less equivalent to the proper adjustment according to U.

MR. STOCK: If you were measuring a number of characteristics, which one would you tie σ to?

DR. NEYMAN: I welcome this question. It is true that we practically never plan an inquiry in order to determine just one single mean. But usually it is not difficult to see that one of them is of greater importance than the others. If such be the case, then the numbers of samplings should be adjusted accordingly. Alternatively, if there are several characteristics of equal importance, we may look for one that could be called the basic characteristic, and which would have the property of being correlated with the ones that we are interested in. This correlation may be positive or negative, but the resulting correlation between the corresponding σ_i will always be positive. Therefore an adjustment according to the basic characteristic will always tend to improve the accuracy.

DR. WILCOX: If you had been advising the Italian census people, what specific advice would you have given?

DR. NEYMAN: I would have advised them to consider their circondarî not as units of sampling but as strata. These strata should be subdivided into units of sampling as small as the character of the material would permit--parishes, streets, single houses; whatever is 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 be more satisfactory if instead of sampling circondarî they would sample parishes. In this of course they are perfectly correct. The results would have been much better yet if they had sampled proportionately to the sizes of the strata.

There is a special difficulty in carrying out an inquiry based on a random sample, which seems to be worth while mentioning. This is

of a psychological nature. Generally we do not rely on random sampling. Intuitively we are inclined to think that it is not wise to rely on chance, when there is some knowledge available that might guide our steps. I have seen many instances where a feeling of a similar kind has made it difficult to reach a decision on how an inquiry should be carried out. I remember very well the doubts that I had myself. "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 it and see. So 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 something like 1000 sheets from census data or the like, 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 a few warnings.

(a) The population in this experimental sampling must be sufficiently heterogeneous, like the populations that we study in practice.

(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 will sometimes ignore knowledge of certain principles. But it must be remembered that in following the indications of the theory you will make use of some other kind of knowledge, that of mathematical statistics.

DR. WILCOX: I have read that article of Dr. Neyman's (footnote page 90) and I noticed that he spoke of his work in connection with the Polish census at the same time that he was commenting on the work in Italy on the Italian census; therefore I asked him the question if he had been the adviser for the Italian census, in view of the fact that their cards were stacked, and of the difficulty of rearranging them, what advice he would have given. 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 take every tenth or twentieth on down the list and make up the sample that way instead of setting up a set of random numbers or drawing numbers at random and selecting the sample according to that little model or game of chance. Are there any advantages or disadvantages that one should bear in mind in making use of the device of taking every tenth name on the list, every tenth family, house or district?

DR. NEYMAN: I think there is a definite advantage of using a mechanical process of random sampling throughout; that is to say, not

taking every tenth unit as listed. Sometimes it will not improve anything, and your tenth or twentieth house will be as good. But there is just the possibility, especially in new properly planned towns, that if you start with every twentieth or fifteenth house you will be synchronized with something very essential in the town itself. We know of one small inquiry where they took a sample of houses in a few villages. The houses were numbered and they decided to take every fifth or every tenth, perhaps, and they obtained something very surprising. Eventually they found out that the first house always was the one belonging to the squire. In new towns you can expect that every block will have the same number of houses, and if you take every fifth house, you may either omit corners or systematically include all of them in the sample, and this may introduce a considerable bias.

It is essential to be clear about the exact nature of the procedure suggested. It is this. We take ten first units of sampling as 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, ..., etc. It will be seen that this procedure is equivalent to a division of the population sampled into 10 parts, thus:

1st part,	sampling units	No.	1, 11, 21, 31, ...
2d	"	"	"	2, 12, 22, 32, ...
3d	"	"	"	3, 13, 23, 33, ...
.				
.				
.				
10th	"	"	"	10, 20, 30, 40, ...

Next we treat those 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 only on good luck, hoping that the ten parts into which the whole population is divided are very similar from one to another. I would recommend that one rely on chance as governed by the empirical law of big numbers, but I would 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* which may be recommended for the purpose. If your sampling units are listed and numbered, 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. You ignore also all numbers exceeding the total of your sampling units.

---

* See footnote on page 14.

DR. LANG: I don't see how this system should be applied concerning names that are listed alphabetically.

DR. NEYMAN: 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 house will give as good a sample as the 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 the attempts to "cut a hair in two along its length." Here the research worker must acquire some experience and use his own judgment in every practical case. It must be emphasized, however, that the use of Tippett's numbers does not present any difficulty at all, and that using them you are on the safe side.

MR. WERTHEIMER: In Dr. Gini's work did he usually make only the mean conform?

DR. NEYMAN: He used 12 controls.

MR. WERTHEIMER: What would he use to make the samples, would he use the averages, the average only?

DR. NEYMAN: The averages of 12 controls.

QUESTION: Was it ever tried making two characters the same?

DR. NEYMAN: Yes, by Professor Anderson in Bulgaria.* They sampled villages. From a previous census they got distributions of various characters of farms within the villages. For each village they constructed the histograms and then they tried to select such villages to form the sample for which the histograms of several controls were similar to those for the whole population. I think it is again a faulty method, but there is no evidence of what happened. I don't know whether any comparison between the characters of the sample that were not used for its selection and the corresponding ones of the population was ever published.

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 having the different areas is that there may be economic factors that affect the unemployment rate where there is a small part of the industry, contrasted with the case 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?

---

* Oskar N. Anderson, Einführung in die Mathematische Statistik (Julius Springer, 1935) p.291.

DR. NEYMAN: 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 a similar equality in accuracy for all industries. Your situation is more complicated than those considered before. The different areas you mention must be considered as separate populations--let us call them partial populations. They may be and should be stratified. Denote by $m_o(i)$ the total number of sampling units to be selected from the ith partial population. This number should be distributed among the strata according to the rule I have indicated before (page 96). If this is done, then the variance of a mean like the one in formula (5) page 96, but now referring to the ith partial population, will be, owing to formula (7),

$$N(i)[\sigma(i)]^2 = \frac{M_o(i)-m_o(i)}{m_o(i)} \sum_{j=1}^{s(i)} M_j(i)[\sigma(i)]^2$$

$$- \frac{M_o(i)}{m_o(i)} \sum_{j=1}^{s(i)} M_j(i)[\sigma_j(i) - \frac{1}{M_o(i)} \sum_{j=1}^{s(i)} M_j(i)\sigma_j(i)]^2$$

$$= \frac{1}{m_o(i)} \{[\sum_{j=1}^{s(i)} M_j(i)\sigma_j(i)]^2 - \sum_{j=1}^{s(i)} M_j(i)[\sigma_j(i)]^2\} \qquad (7a)$$

where the notation of Eq.(7) page 96 has been altered so that the i in parenthesis refers to the ith partial population, and the subscript j to the jth stratum. If, as formerly, we denote by m_o the total number of sampling units to be selected from all the partial populations, then this number of samplings must be distributed between those populations so as to keep $\sigma(i)$ in formula (7a) constant. This, of course, refers to one industry only, and assumes knowledge of the $\sigma_j(i)$ for each stratum and for each partial population. The values of $\sigma_j(i)$ could be estimated from a preliminary inquiry.

MR. KANTOR: In attempting to get an estimate of the variability that we are going to use in deciding the proportion that you will draw, you will have to take a test count in each of your areas; you have it scattered over a number of characteristics, it is no longer one characteristic that you measure. You would have to get a test drawing and compute actual unemployment rates for a number of industries in each of your areas. 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.

DR. NEYMAN: I don't think so. The preliminary inquiry designed to estimate the variability of the strata may be very small in size. As I have already mentioned (page 98), Dr. Sukhatme has investigated this question and found that 20-30 units of sampling out of each stratum would be plenty. He suggests even as few as 15. And it is not necessary to make a preliminary inquiry separately for each industry. You make one and use it to estimate $\sigma_j(i)$, for each of the industries in turn. Substitute your estimates for the true $\sigma_j(i)$ in formulas (4) and (7a), separately for each industry. You will see that formula (4), indicating the

optimum proportions of samplings within strata, will give more or less similar results for all industries. Probably--but of this I am less certain--the same thing will happen with formula (7a). Alternatively you may adjust your proportions of sampling to some single character treated as basic. I should choose the total number of workers within the sampling unit; it is likely to be highly correlated with the numbers of unemployed.

MR. KANTOR: We find, however, in industry that there are very great differences in the proportion unemployed, depending on the production rate of the industry to which the workers were 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 largely devoted to the latter type of production will have small unemployment. Is it that variability that we can test by drawing a small preliminary sample?

DR. NEYMAN: Yes, that is quite all right. The variability you speak of does not cause any trouble since this is a variability between the strata, or perhaps between the 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. My impression even is that it is partly the purpose of your study to get information concerning such areas separately. 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 n, for the moment, the number of workers within a unit of sampling, and by x the number of them previously working in some particular industry and now unemployed. If you take one particular stratum and study the units of sampling, you may find a picture something like this:

Values of n	100	150	35	10	200	etc.
Values of x	10	13	1	2	25	

We have seen something of that sort in an actual inquiry in Poland. The total number of workers within the boundary of a unit of sampling is composed of various kinds of these workers, and n is bound to be correlated with x. Of course this correlation is merely due to the varying size of the sampling unit.

The plan to take but one basic character as a unit has the advantage that in using it you can tell if the preliminary inquiry might be made very superficial and yet be satisfactory; the enumerators might then be asked to establish only the number of workers inhabiting the units of sampling. But this procedure has also definite

disadvantages. First of all if you work only with the basic character, the data collected during the preliminary inquiry could not be included into the main one. Next the basic character is very likely to be useful for assigning the numbers of sampling to separate strata of one partial population, but I am not so certain whether this will be the case when you try to determine the level of sampling of partial populations. Therefore I should probably carry out the preliminary inquiry exactly as the main one, with the only difference in size. I would estimate each $\sigma(i)$ separately for each industry and substitute it into formulas (4a) and (7a). 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 the $m(i)$.

MR. 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 the information on expenditures.

The question I should like to ask is whether 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?

DR. NEYMAN: So far as I know, nothing has been done on the specific question that 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 would, as I have already indicated,

need to cover only a relatively small number of cases.

MR. FRIEDMAN'S QUESTION RESTATED BY DR. WILCOX: For part of the work at least one step was taken in trying to get a random sample using every nth card, starting not with the first card but with the 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. That was an inexpensive survey. Then the long laborious process had to be followed of finding out how they spent their money in detail, and the number of families that might respond 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 around. What should be the relative number that would 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 to the smaller?

DR. NEYMAN: I repeat, so 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 it could be answered by some simple method. If I succeed I will certainly try to publish the results.*

. . . . . .

The other group of questions that I was asked to answer refers to standard economic problems. The persons who asked those questions were interested to know how far statistical methods could be used to illustrate some connections between different economic factors developing in time. We know that there are cycles. We know that there are inter-relations between movements of prices, of unemployed, and so on. All those factors are developing in time, and if we plot the figures, we observe some parallel movements, or movements in opposite directions, and it is a question how far statistical technic and in particular the theory of testing hypotheses, or the theory of sampling, is applicable to such observations.

---

* Editor's note: During the summer of 1937, Dr. Neyman prepared a paper dealing with the question raised by Mr. Friedman and Dr. Wilcox, and it is slated to appear in the March 1938 issue of the Journal of the American Statistical Association. A paper by Milton Friedman, reporting the survey of family expenditures, along with a new method of handling ranked data, will appear in the issue for December 1937, under the title of "The use of ranks to avoid the assumption of normality implicit in the analysis of variance."

My opinion is that the phenomena described could and should be studied statistically, but that the appropriate methods have not yet been properly worked out. The procedure that is ordinarily applied seems to be wrong. Many people feel this sub-consciously, and this is probably the reason why similar questions are being asked so frequently.

Let us make a short review of the methods that have been applied in studying the correlations between two time series.

We start by trying to split each of the series into several parts, which we arbitrarily assume to be additive. One of these parts is the trend, which we estimate perhaps by fitting a low order parabola to the whole series available. The next part is the "business cycle." The third part is the "seasonal variation," which we frequently estimate by calculating moving averages. Finally, the remainder is considered to arise from random causes, and we concentrate on the question whether such a remainder in one of the variables is correlated with that in some other.

All this procedure seems to me very artificial and arbitrary. How do we know that the trend, the business cycle, and the rest, are connected together additively? Why should all the systematic variation, everything except the random component, be represented by smooth curves? Finally, even if all the hypotheses described were true, we must notice that the residuals calculated in the above manner are not equivalent to direct observations but are in some way or other related. Consequently, if they are used for testing hypotheses, some novel methods, not yet available, must be devised. This, however, does not seem to be a fruitful field of research.

In my opinion the whole problem of time series must be treated from a point of view that is quite different from the traditional one just described. As a matter of fact, we are already witnessing some attempts on these new lines. Work is being carried on in two directions. One direction is represented by the authors who try to formulate intelligible hypotheses concerning the machinery of economic phenomena and to express them in terms of either differential equations or equations in finite differences. Solving those equations they obtain measures of various interacting economic factors as functions of time. The corresponding curves (see, for instance, diagram I page 116, in the conference on time series), resemble in some respects the ones that may be given by observation, but there is one essential characteristic in which there is a distinction between that theory and the observations: the theoretical curves are too regular. And this is just the circumstance indicating the need for a new branch of mathematical statistics that must be developed for treating time series. The excessive regularity in the theoretical curves is an obvious result of the circumstance that there is nothing variable in conditions from which they are deduced. Now in actual life, the rigid economic hypotheses that have led to the ordinary differential equations may be satisfied approximately but not

exactly. It may be, for example, that the purchases of the rural population in each year are, as a rule, nearly proportional to the consumption of agricultural produces in the preceding year; but it is obviously impossible to expect that the coefficient of proportionality will be an absolute constant.

It follows that if we desire a better agreement between observations and the dynamic theory of economics, the theory must be based on differential equations or equations in finite differences of a special kind: the coefficients in these equations, at least some of them, must not be rigid constants but random variables. The solutions of such a system of equations would not be represented by definite curves, even if the initial conditions were fixed. Instead, for any fixed moment t, we should have a probability distribution of the variables concerned, depending on the values that they had in previous moments. Having a system of dynamic hypotheses, the solution of the random differential equations resulting from them, and the observational data, we could test whether the agreement is satisfactory. This is, in my opinion, the right way of treating time series.

Unfortunately the theory of random equations is not yet ready or, more precisely, it is only just being started; and this is the other direction in which the theoretical work concerning time series is being carried on.

MR. STOCK: Would you give a reference on this last subject you have been talking of?

DR. NEYMAN: To my knowledge, the first problem of this kind was treated by Professor Hotelling.* It consisted in finding the most probable size of a population at a moment between two censuses. Seeing that the most difficult problem in scientific research consists in noticing that there is a problem at all and in formulating it properly, I think that this paper is a very brilliant achievement of its author.

The theory of random equations has been discussed by S. Bernstein* in his report to the International Mathematical Congress in Zurich 1932, and is a series of papers published in the Proceedings of the Leningrad Academy. Some of them are in French, others in Russian with extensive French summaries.

. . . . . .

---

* The reader is now referred to the conference on time series. The reference to Hotelling's and Bernstein's papers will be found on page 119.

MR. STOCK: In regard to the first series of questions, is there any way of estimating the effective size of the sample in order that we may estimate a standard error from these stratified samples? If we have in our sample a thousand persons, but we have sampled only 500 houses, what is the effective size of that?

DR. NEYMAN: The size of the sample is not the number of persons, but the number of random selections made to form the sample. Therefore, if you select houses at random so that your unit of sampling is a house, then it is the number of houses that is relevant. If towns are selected, only 25, and those 25 towns contain say 25,000,000 people, then the size of the sample would be 25, not 25,000,000.

In order to estimate the effective size of the sample required to get a given accuracy of the estimated means you may apply the formula for the variance σ^2 of X as given in Eq.(7) page 96. Of course you will want the estimates of σ_i; those could be found from the preliminary inquiry,

TIME SERIES ANALYSIS AND SOME RELATED STATISTICAL PROBLEMS IN ECONOMICS

A conference with Dr. Neyman in the auditorium of the Department of Agriculture, 10th April 1937, 11 a.m., Dr. Charles F. Sarle presiding.

I shall speak of two different ways of treating economic problems by means of statistical methods. One could be called empirical, another a priori. The first is very popular, has yielded many important results in the past, and is likely to be useful for a considerable time in the future. The other method is much less popular and so far has not proved very efficient. Still, I shall rather condemn the empirical method and praise the a priori. Both methods are designed for the same purpose: to make predictions concerning economic processes as described by various figures such as prices, incomes, supply, and demand. Their distinctive characteristics are as follows.

With the empirical method the dominant hypotheses concern the final result of the work of the economic machinery in any given situation, without paying attention to the machinery itself. On the other hand, treating the problem a priori, we start by formulating some hypotheses about the machinery, and then work out the results that can be produced from those hypotheses.

Before giving examples taken from the current literature and criticising the two methods of approach, I shall try to emphasize the distinction between them, by quoting certain instances from the history of astronomy, which, you know, could be roughly divided into two phases, before Newton and after Newton, with a transitional period marked by the names of Copernicus and Kepler.

All, or almost all, the work done in the first phase was independent of any hypothesis concerning the machinery of movements of celestial bodies. On the other hand the authors firmly believed that, whatever this machinery may be, it must have produced only circular movements. The circles could have been stationary with their centres at the Earth, or they could have been moving; but then their centres should have moved along other circles, the final circle being centred at the Earth. The astronomers of this phase aimed at the establishment of a complex system of circular motions that would agree with the observations. The predictions based on such systems were frequently very successful.

After the works of Copernicus, abolishing the dogma that the centre of the last circle must be at the Earth, and after those of Kepler, dealing similarly with the other dogma of circularity, the way for Newton's theory of gravitation was cleared and modern astronomy was begun. This is distinguished by the circumstance that, instead of making assumptions concerning what must be the nature of the observable movements, it formulates hypotheses concerning the machinery that may have produced these motions. The deductions from these hypotheses were then compared with observation. This is what I have called the a priori method. You know that in the case of astronomy, this proved to be much more efficient than the empirical method. I am induced to think that something similar will prove to be true in economics.

Empirical statistical research in economics may be exemplified by a recent work of E. C. Rhodes.*

Various institutions in different countries are engaged in computing and in publishing indices measuring the extent of various economic activities in terms of some conventional unit, usually representing the same measure at some fixed moment. Apart from such indices, like those of building activity, circulation of currency, electric power, production of various commodities -- indices which I shall call specific--many economists find it necessary to consider something that could be termed the index of general business activity. The object of Rhodes' paper was to supply a new method of determining such an index of business activity and he used it to calculate the values of that index for Great Britain.

His way of reasoning was approximately as follows: Consider a number of specific indices and their values as calculated for moments $t_1, t_2, \ldots, t_n$. Let those specific indices be $X_1, X_2, \ldots, X_s$. The changes in the values of those indices occurring in time depend presumably on a very great number of economic factors, and Dr. Rhodes assumes the existence of one particular factor, supposed to influence all of the indices, which he calls the index of business activity, denoted by I.

He does not specify any properties of this factor, that is to say he does not define its nature. Nor does he make any hypothesis concerning the machinery of economic processes involving the influence of I on the X_i, but he does assume a certain particular form of dependence between the X_i and I. In other words, he makes the assumption that whatever the machinery of economic processes may be, the final results of this machinery must express itself by equations of a particular kind connecting the X_i with I, and also some other hypothetical and equally undefined factors.

You will have no difficulty in recognizing here the same way of approaching the problem as we have seen in the first phase of the history of astronomy.

The method of Dr. Rhodes is, I dare say, familiar; and probably many of you would easily guess the nature of his equations connecting the X_i with the mysterious factors. They are linear equations with some unknown constant coefficients. Denoting severally the unknown economic factors influencing the X_i by ξ with subscripts, we may write the equations assumed by Dr. Rhodes in the form

$$X_i = a_i I + b_{i1}\xi_1 + b_{i2}\xi_2 + \ldots + b_{im}\xi_m \quad \text{for } i=1, 2, \ldots, s. \quad (1)$$

The values of I and the ξ's are changing with time, and it is assumed that they do so independently of each other; and the values of each of the X_i are supposed to change in accordance with the above equation. In the analogy with astronomy, this assumption corresponds to the motion of a celestial body, which motion may be split into a number of cyclical motions. But there is another assumption that Dr. Rhodes makes which could be considered as roughly corresponding to that of the last of the Ptolemy's circles centring at the

---

* E. C. Rhodes: "The construction of an index of business activity", Journal Royal Statistical Society 100, pp. 18-66, 1937.

Earth; this is that making some linear combinations of the equations (1), we can manage to eliminate most of the ξ_i and so obtain the so called subsidiary equations of the following form:

$$\begin{aligned} A_{11} X_1 + A_{12} X_2 + \ldots + A_{1s} X_s &= B_1 I + C_1 \xi_1 \\ A_{21} X_1 + A_{22} X_2 + \ldots + A_{2s} X_s &= B_2 I + C_2 \xi_2 \\ &\vdots \\ A_{s1} X_1 + A_{s2} X_2 + \ldots + A_{ss} X_s &= B_s I + C_s \xi_s \end{aligned} \qquad (2)$$

The final formula for the calculation of approximate values of I, for each moment for which the values of the X_i are available, represents a weighted average of the X_i, the weights depending in a certain manner on the coefficients of the subsidiary equations (2). Consequently the possibility of calculating the values of I depends on that of estimating the coefficients in (2). Dr. Rhodes gives a way of estimating them. Of this I shall notice only that it is based solely on the values of variances and product moments of the X_i. That is to say: if we use the observed values of the X_i, say

$$X_i(1), X_i(2), \ldots, X_i(n) \qquad i=1, 2, \ldots, s. \qquad (3)$$

where $X_i(t)$ denotes the value of X_i observed at the moment t, to calculate*

$$\overline{X}_i = \frac{1}{n} \Sigma X_i(t) \quad i=1, 2, \ldots, s. \qquad (4)$$

$$s_i^2 = \frac{1}{n} \Sigma [X_i(t) - \overline{X}_i]^2 \qquad (5)$$

$$p_{ij} = \frac{1}{n} \Sigma [X_i(t) - \overline{X}_i][X_j(t) - \overline{X}_j] \qquad (6)$$

then all the calculations suggested by Dr. Rhodes, perhaps in a different form, could be carried out to obtain estimates of the coefficients in (2), these estimates depending only on the values of the s_i and p_{ij}.

The method of Dr. Rhodes has the advantage that it cannot fail to produce a result, that is to say, some result. Similarly, the efforts of early astronomers always produced a system of circles and velocities, bringing their theory into fair agreement with the observed motions of the stars and planets. And the objections to both methods are similar. They are summed up in the question of the reality of the objects with which the respective theories are dealing. Do the epicycles actually exist? Is there any reality corresponding to Dr. Rhodes' index of business activity and other factors?

---

* Unless otherwise indicated, the summation Σ will be taken over t running from 1 to n.

"Reality" is a dangerous conception. Do atoms represent a reality? This may be questioned. But whether they are or not, the chemist treats them as such; in other words, he postulates their reality. Therefore, I shall formulate the two above questions a little differently: is there any advantage in postulating the reality of the epicycles in one case and of the various "economic factors"--such as, for example, the general business activity--in the other?

We are driven here to a more general question, what advantages we might expect at all from postulating the reality of this or that conception? But this question does not seem to be very difficult. Its answer is readily obtained when we think of conceptions that are strongly established in science and in the reality of which we have forgotten to doubt. The advantage of postulating the reality of atoms in general and of the atom of hydrogen in particular consists in the fact that their definition--in the textbooks of chemistry treated as <u>discovered</u> properties--permits us (i) to identify the particular atoms in various conditions; and (ii) to predict the results of various experiments. It was possible to bring the definition to such a remarkable concordance with the facts of observation that various complicated checks and rechecks are matters of history.

We may start with a quantity of matter, conforming with the definition of hydrogen, bring it into contact with oxygen, and produce a quite new kind of matter, water. This again could be transformed into a number of other substances. As then the atoms of hydrogen do not lose their identity, they can be obtained again in their previous state, and, apart from an experimental error in the previous quantity.

Similarly there is an advantage of postulating the reality of forces in general and that of gravitation in particular. Here again we have the possibility of identification of generality and of relatively easy predictions in an enormous number of cases, including movements of planets.

If, armed with such observations, we turn to the problem treated by Dr. Rhodes, or indeed to any other problem treated empirically, we shall see that there are no similar advantages in postulating the realities that are usually postulated in empirical methods. If there are any advantages at all, they must be different. But I cannot perceive them.

If we do not build any system of hypotheses concerning the machinery of processes under discussion, it is <u>eo ipso</u> difficult to give definitions to the assumed realities that will permit one to distinguish them from others. It follows that, as a general rule, in any two given cases, there may be difficulties with identification of the same realities. And I cannot see any use in science of "realities" that are impossible to identify. To illustrate my point I shall assume for a moment that the economic factors considered by Rhodes are real and that in some three different countries A, B, and C a certain set of three indices X_1, X_2, and X_3 are connected with factors ξ by the following equations:

Country A

$$\left.\begin{aligned} X_1 &= I + \xi_1 \\ X_2 &= I + \xi_2 \\ X_3 &= I + \xi_3 \end{aligned}\right\} \qquad (7)$$

Country B

$$\left.\begin{aligned} X_1 &= I\,\tfrac{1}{2}\sqrt{3} + \xi_1\,\tfrac{1}{2}\sqrt{5} \\ X_2 &= I(1/2\sqrt{3}) + \xi_1(3/2\sqrt{5}) + \xi_2\sqrt{22/15} \\ X_3 &= I(2/\sqrt{3}) + \xi_2\sqrt{10/33} + \xi_3(2/\sqrt{11}) \end{aligned}\right\} \qquad (8)$$

Country C

$$\left.\begin{aligned} X_1 &= \xi_1 + \xi_2 \\ X_2 &= \xi_1 + \xi_3 \\ X_3 &= \xi_2 + \xi_3 \end{aligned}\right\} \qquad (9)$$

You will see that the situation in the three countries is entirely different and that in country C there is no general factor I at all. Yet, if all the factors, I and the ξ_i, are mutually independent and varying about zero with standard deviations equal to unity, then, as it is easy to calculate, the variances of all the X and also their product moments will be in all the three countries identical, namely

$$s_i^2 = 2,\ p_{ij} = 1,\ i \neq j;\ i,j = 1, 2, 3. \qquad (10)$$

Thus, the method of Dr. Rhodes, if applied to data from any of the three countries, is bound to give the same results--that is, apart from some fluctuations due to random variation. Moreover, if all the variables, I and the ξ_i besides being independent, happen to be normally distributed, then the distributions that the X_i will follow in countries A, B, and C, will be completely identical in all respects, and not only in their variances and product moments. It follows that in such a case, not only the method produced by Rhodes, but also any other imaginable method, would not enable one to distinguish between the situations as described by (7), (8), and (9). Yet, they are greatly different, and if the equations (7) and (8) were known, they would yield different formulas by calculating I, namely

For country A: $I = \text{const.}\ (X_1 + X_2 + X_3)$ (11)

For country B: $I = \text{const.}\ (X_1 - X_2 + 2X_3)$ (12)

For country C, similar calculations would be useless, since no common factor exists. However, Dr. Rhodes' method would provide means for its approximate calculation --the same formulas as for countries A and B!

Seeing all this one is led to ask: What is gained by the assumption of the reality of various unlefined factors, and what is gained by the methods of their calculation?

The situation would be quite different if the factors were defined in a way permitting their identification and a more or less direct measurement. Even if the distributions of the X_i in my three countries were identical, there may have been a possibility of distinguishing between the prevailing situations on some other grounds. But then the method of approaching the problem would not be what I call an empirical one. *

I have dwelt on the work of Dr. Rhodes partly because it is such a good example of a purely empirical method of approach and partly because it was published only recently. I remember it in detail. But it is one of many similar ones. If you consider the method of the so-called confluence analysis, advanced by Ragnar Frisch, you will find that it is open to almost identical criticism. Quite similar objections apply also to most of the popular methods of dealing with time series. We usually make arbitrary assumptions concerning not the nature of the economic processes but the nature of the functions of time representing their results. We say, namely, that the time series that we observe are sums of four independent components: the secular trend, the business cycle, the seasonal variation, and the random residual. Next we postulate that certain algorithms are able to eliminate one or more of these components and so to estimate the others. We apply the algorithms and obtain some results, but I do not think that they will help much in the process of building a new theoretical economics, in the sense of the words in which we speak of theoretical astronomy.

All that these results can give is comparable to what gave the works of the early astronomers a multifold but unsystematic knowledge of many unconnected facts. This, of course, is very valuable.

Let us now turn to the other kind of statistical research in economics. I hesitated a little before the adjective statistical as used just now; perhaps it would be more proper to say mathematical, but I think that, if not the present, then the future will justify statistical. At present, however, there is not much statistics in the work. It consists chiefly in putting down a few hypotheses of the kind that we frequently see in analytical mechanics, concerning certain economic magnitudes and expressible by means of equations either in differentials or in finite differences, and in deducing consequences of the same.

A good example of this kind of research is provided in the first part of

---

* I want to emphasize that the above criticism of the empirical method of approach illustrated in Dr. Rhodes' treatment of the problem of the index of business activity, does not imply in any way that, in my opinion, the calculation of such an index is useless. As a conventional measure of changes occurring in a number of particular economic activities it is probably useful. But to be useful, it must be clearly defined in one way or another, as for example, in the case of an index published in the "Economist": a certain weighted average of a number of specific indices with some fixed weights. What I object to is the postulation of some really existent undefined and unidentifiable factors, among them business activity, the postulation of the form of the equations connecting them with the observable indices, and the attempt to measure what is not defined and impossible to identify.

a paper by Ragnar Frisch,* in which the author formulates certain simple hypotheses about the exchange of products between the urban and rural populations. The former is personified by a shoemaker, the latter by a farmer. As a first stage in the study it is assumed that the amounts of money x_t and y_t spent during the t-th year by the shoemaker on farm products, and by the farmer on shoes, are proportional to their own sales during the preceding year, so that

$$\left.\begin{aligned} x_t &= a y_{t-1} \\ y_t &= b x_{t-1} \end{aligned}\right\} \qquad (13)$$

where a and b are some positive constants.

Frisch solves those equations and discusses all the possible consequences that they may involve. Later on he mentions that the original scheme of the exchange is probably too simple and he makes an additional hypothesis, expressible by saying that the purchases of each of the two parties are influenced by their mutual indebtedness. Denote by

$$G_{t-1} = \sum_{i=1}^{t-1} (y_i - x_i) \qquad (14)$$

the amount of money owed by the farmer to the shoemaker at the beginning of the t-th year. Then it is assumed that

$$\left.\begin{aligned} x_t &= a y_{t-1} + c\, G_{t-1} \\ y_t &= b x_{t-1} - d\, G_{t-1} \end{aligned}\right\} \qquad (15)$$

These new equations are solved and the implications discussed. I am not an economist and it is difficult for me to judge how far those particular hypotheses are likely to cover the essential part of the process of exchange between the town and the village. I suspect that the situation is oversimplified. But, however that may be, it is my opinion that, if a satisfactory general economic theory will ever be built at all, then it will emerge out of such oversimplifying hypotheses.

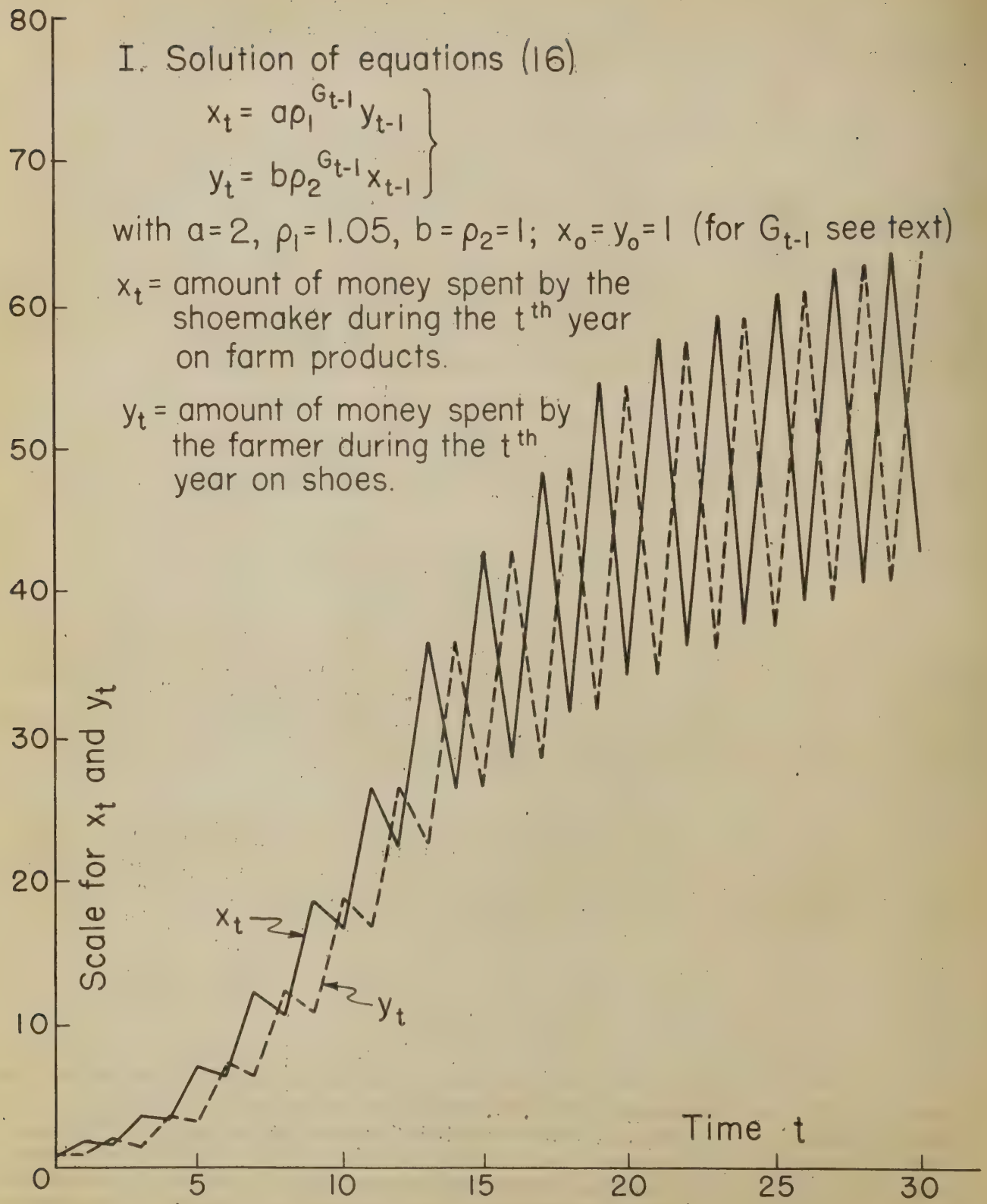
I must, however, mention a defect in the above treatment. This will be apparent if we consider the curves representing the year-to-year variations of x_t and y_t as determined either by (13) or by (15), or by any other system of ordinary equations. Diagram I on the next page gives the graphs of x_t and y_t forming a solution of the following system:

$$\left.\begin{aligned} x_t &= a\rho_1^{\,G_{t-1}}\, y_{t-1} \\ y_t &= b\rho_2^{\,-G_{t-1}}\, x_{t-1} \end{aligned}\right\} \qquad (16)$$

with $b = \rho_2 = 1$, $a = 2$, $\rho_1 = 1.05$, and $x_o = y_o = 1$. I do not know whether the system (16) has any advantage over (15) from the economic point of view, but I prefer it because it cannot lead to negative values of x_t and y_t. Looking at the diagram (p.116) we see at once that it cannot represent the movements of any sort of living business. The curves are too stiff. The zigzags may perhaps be mistaken for what in ordinary time series is treated as "random residuals," but only at first sight. They are distinctly too regular for accidental occurrence.

---

* Ragnar Frisch, "Circulation planning:...," Econometrica 2, 258-336, 1934.



The coordinates of the points will be found on page 125.

It is obvious that whatever system of ordinary equations in finite differences we take, the curves so determined will all be of the same kind--they will be too regular to represent the movements in economic life. It follows that ordinary equations in finite differences, and all integral and differential equations, are not the proper tool for dealing with economic problems. We must invent something else.

To describe what I think is the proper tool I have prepared diagram II (next page), representing the variations in x_t and y_t, the yearly purchases of the shoemaker and the farmer from each other. It is easily seen that the curves in this second diagram differ essentially from the curves in the former one (p.116) in that they are not "dead". If the former were shown to any statistician he would not hesitate to state that it could not represent any real economic process, whatever be its kind. This is not true with respect to curves in diagram II, as many actually observed time series look very similar: you may distinguish here something like an ascending trend, two cycles; and also distinctly random fluctuations. Therefore, if any mathematical theory could have produced those curves, then we should not reject it without further investigation. At least we should not reject it for the same reasons that cast doubt on whether the ordinary equations in finite differences, such as (15) or (16), could be used directly to represent the machinery of the real economic processes.

As a matter of fact, the curves in diagram II did come from some mathematical scheme, and this differs but little from the system of equations (16).

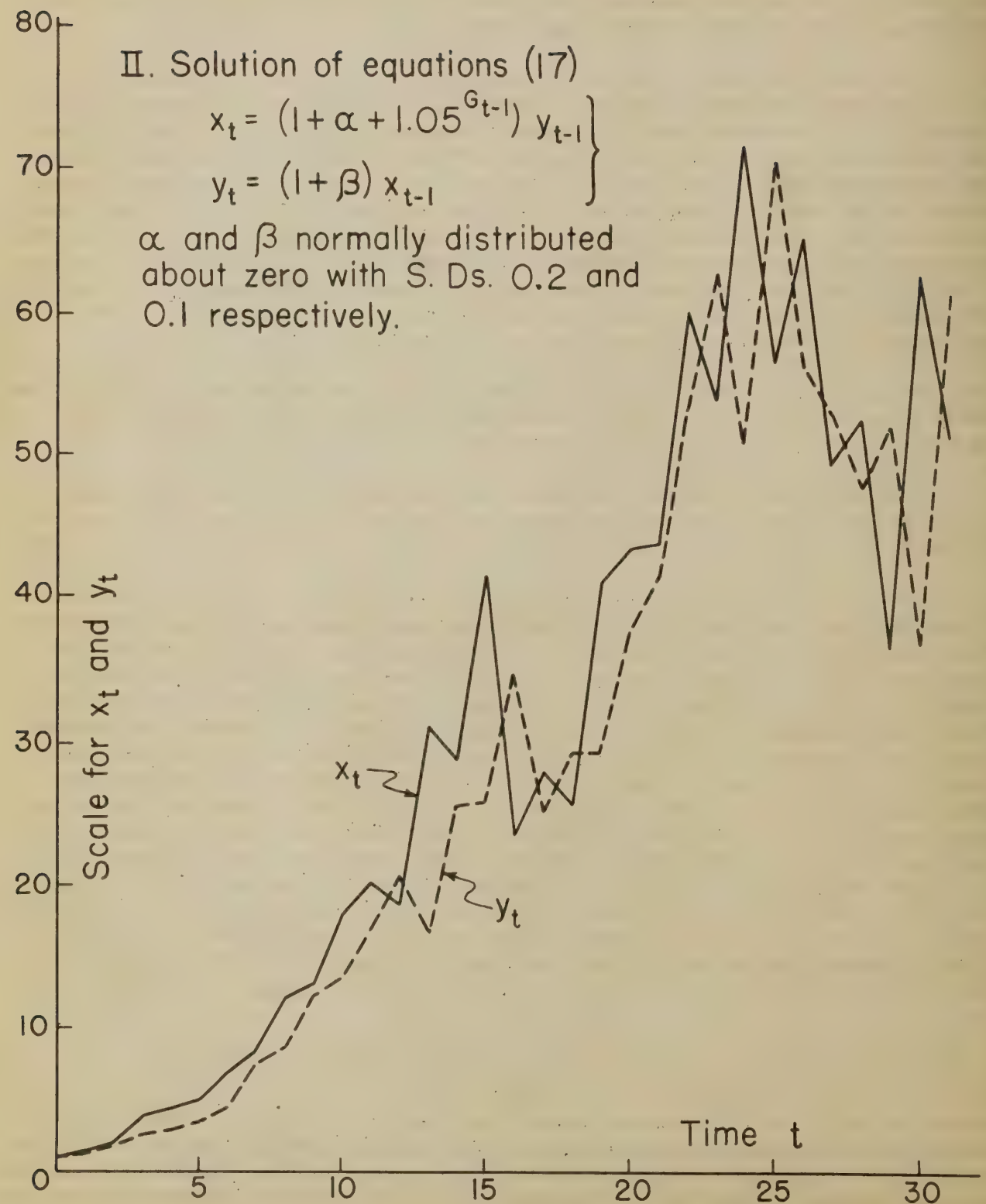
We may notice by just looking at them more carefully that the equations (16) are not likely to represent any real process. Assuming even the very simplified situation where x_t and y_t represent the yearly mutual purchases of only one shoemaker and only one farmer, respectively, we should hardly believe that they could adjust their purchases so as to satisfy equations (16) exactly. It may be an <u>intended policy</u> of the farmer to purchase from the shoemaker exactly as many shoes (in monetary units) as he sells him food, but it is almost certain that in practice there will be deviations from this intended policy. The same is true with the shoemaker. Therefore, equations like (16) may be considered as expressing the general tendency of the parties; but if they are intended to represent what really happens in practice, they must be modified to include a random element, which is the necessary attribute of anything concerning living societies. The symbols that in these equations represent the constants --some of them at least -- must represent random variables. And this is exactly the way in which the curves of diagram II were obtained. It was in fact assumed there that

$$\left.\begin{aligned} x_t &= (1 + \alpha + 1.05^{G_{t-1}})\, y_{t-1} \\ y_t &= (1 + \beta)\, x_{t-1} \end{aligned}\right\} \quad (17)$$

where α and β are independent variables, each following a normal law about zero with S.Ds. $\sigma_1 = 0.2$ and $\sigma_2 = 0.1$ respectively. The values at the origins were assumed to be as formerly, $x_0 = y_0 = 1$. The technique of obtaining the curves of diagram II consists in the following: We start by reading from tables of normal deviates * a number of figures that might have been the

---

* See the footnote on page 79 referring to the tables of Mahalanobis.



The coordinates of the points will be found on page 125.

independently observed values of a normal variable, varying about zero with S.D. equal to unity. Dividing the first 31 of these numbers by 5 (since $\sigma_1 = 0.2$) and the following 31 by 10 (since $\sigma_1 = 0.1$), we obtain what might have been the values of α and β respectively. Substituting those values in turn in (17), the consecutive values of $x_1, y_1; x_2, y_2, \ldots, x_{31}, y_{31}$, were obtained, and those give the curves of diagram II. A curve of this kind was obtained by Ragnar Frisch (loc. cit., p.271). It corresponds to the system of equations (15) subjected to "erratic shocks." These "shocks" certainly mean a random variation of some of the coefficients in the equations, but the details are not given by the author.

It seems to me that the proper way of approaching economic problems mathematically is by equations of the above type, in finite or infinitesimal differences, with coefficients that are not constants, but random variables; or what is called random or stochastic equations. They can embody the a priori hypotheses covering the machinery of the economic processes studied; that is to say, hypotheses like those of Ragnar Frisch just described; and then they will leave room for chance variation, which seems to be an essential feature of any real time series.

It must be noticed, however, that this tool is not yet quite ready for use in economic studies. The theory of random differential and other equations, and the theory of random curves, are just starting their existence. The number of papers published on the subject is small.*

We must understand what a solution of a random equation in finite differences may represent. For simplicity I shall consider only the case of two time series, x_t and y_t, as in the graphs just discussed. Obviously the solution of a system of two random equations cannot give a unique system of values of x_t and y_t for each moment t, x_o and y_o being fixed; and in this way the random equations would differ from any ordinary difference equations. Instead of a unique set of values for x_t and y_t we have a probability distribution of them, generally depending on the values $x_o, x_1, \ldots, x_{t-1}, y_o, y_1, \ldots, y_{t-1}$, which the variables x_t and y_t had at previous moments. Having this probability distribution, say $p(x_t, y_t | x_o, \ldots, x_{t-1}, y_o, \ldots, y_{t-1})$, we could calculate the most probable values of x_t and y_t; and this, in certain cases, may be an interesting result. In fact, this would be the form of prediction of future values of the two variables. The product

$$\prod_{i=t_o+1}^{t} p(x_i, y_i | x_o, \ldots, x_{i-1}; y_o, \ldots, y_{i-1}) \qquad (18)$$

would give the probability law of all the x_i and y_i, starting with the

---

* The first contribution in which a particular problem was solved seems to be due to Harold Hotelling: "Differential equations subject to error, and population estimates," J.Amer. Stat. Assoc. 22, 283-314, 1927. The first author to discuss the general theory of random differential equations seems to be S. Bernstein, of which I shall cite his report read before the International Mathematical Congress in Zurich, 1932; and a paper by him entitled: "Principes de la théorie des équations différentielles stochastiques" published in the Memoirs of Stekloff Institute, Leningrad Academy, Vol.V, pp.95-124, 1933. Finally I shall mention a paper on the theory of random curves by E. Slutsky: "Qualche proposizione relativa alla teoria delle funzioni aleatoarie" Giornale dell' Istituto Italiano degle Attuari, Anno VIII, pp.183-199, 1937.

(t_o+1)st pair and ending with the t-th. Integrating the result with respect to x_{t_o+k}, y_{t_o+k} for $k = 1, 2, \ldots, t-t_o-1$, we could obtain $p(x_t, y_t | x_o, x_1, \ldots, x_{t_o}; y_o, y_1, \ldots, y_{t_o})$, the probability law of x_t and y_t, depending on the given known values of $x_o, x_1, \ldots, x_{t_o}$, and $y_o, y_1, \ldots, y_{t_o}$, which it would be possible to use for predicting the values of the two variables corresponding to a moment t-to units of time ahead of t_o.

Finally, the above product (18) corresponding to $t_o = 0$ would give the probability law of all the systems of the x_i and y_i, starting with the first and finishing with the t-th. This could be used, for example, for calculating the most probable shape of the curves representing the two time series.

Our present knowledge permits us to perform all those operations for particular problems, but the number of general results concerning them is not large.* Besides, even the solution of particular problems presents great technical difficulties.

All the above steps depend on the assumption that the machinery of a given economic process is known to be expressed by a given system of random equations. In fact, they are examples of deductions that are to be made from such equations. We must now face the problem of how to decide whether any given system of random equations does or does not represent the machinery of a given process that is described by some time series, actually observed or to be observed in the future.

This problem forms a chapter, as yet untouched, in the theory of testing statistical hypotheses. Here again we shall probably have to wait a considerable time till all the tools that are necessary for economic research are ready. To explain the problems that I see in this section of the work, it will be useful to discuss a simple example.

Suppose that we are able to observe the figures x_t and y_t measuring the mutual purchases of the urban and rural populations respectively, and that we wish to test the hypothesis, say H_o that the whole machinery of the economic process could be described by the random equations**

$$\left.\begin{aligned} x_t &= (a_o + \alpha) y_{t-1} \\ y_t &= (b_o + \beta) x_{t-1} \end{aligned}\right\} \qquad (19)$$

where a_o and b_o are some unknown constants and α and β are two random variables, known to be normally distributed about zero with unknown S.Ds. σ_1 and σ_2 respectively. The initial values of the two variables x_o and y_o will be assumed known.

---

* The above problems are closely connected with that of the so-called Markoff chains forming a special branch of the theory of probability. See V. Romanovsky "Recherches sur les chaînes de Markoff", Acta Mathematica, t. 66, pp. 147-251, 1935. See also J.V. Uspensky, Mathematical Probability (McGraw-Hill, 1937) P. 301.

** It must be clearly understood that these equations are considered only as an example and that it is not suggested that they, or any other in this paper, might represent the actual economic process under consideration.

It will be seen that the above hypothesis H_o, or rather the statement of what is considered as known, is a little artificial. In practice we should probably hesitate to assume normality of the variables α and β. However, as I have said, the present state of the theory of testing statistical hypothesis is only that in statu nascendi, and the above assumption is made in order to make the necessary calculations easier for the deduction of the test.

A test of the hypothesis H_o will be conceived as a rule that is, in a sense, similar to the rules found for games of chance. We consider the situation existing before the values of the x_i and y_i are actually observed, and we denote by W all the possible systems of n pairs of such values, for t = 1, 2, ..., n, that may be given by future observations. Denote generally by E any such system, so that W will be the set of all the E. We agree that, as a result of the test, we shall take one of the following steps regarding H_o: we shall either reject it or refrain from doing so (= "accept"it, for short). The process of choosing a test for the hypothesis H_o becomes thus equivalent to the division of the set W into two parts, say w and W-w, with the intention of rejecting H_o whenever the observed values of the x and y fall within w, and of accepting H_o when they fall in W-w. The problem is thus reduced to that of choosing properly the set w, which is designated as the critical region of W. This may be attempted in various ways, but I shall describe briefly only one of them leading to the particular kind of test called unbiased of type B .*

Not being able to say in advance whether the hypothesis tested H_o is right or not, we have to consider the consequences of any particular choice of the critical set w, both when H_o is true and when it is not.

Assume first that H_o is true. Then we shall want the set w to be so chosen that the probability of the values of the x and y to be observed falling within w should be rather small, say P = 0.05, or P = 0.01, or the like. In general, there are considerable difficulties in finding the particular set satisfying this condition, since the hypothesis tested H_o may be true, and then the values of the unspecified constants, a_o, b_o, σ_1, and σ_2 may vary within broad limits, and this will usually influence the probability of E falling within this or that part w of W. The set w that is wanted is what is called "similar" to W with respect to a_o, b_o, σ_1, and σ_2 and having its "size" equal to P, which means that it must possess the property that whatever be the values of the parameters a_o, b_o, σ_1, and σ_2, if H_o is true, the probability of E falling within w has always the same value, namely P.

In the particular case under consideration there is an infinity of sets w satisfying this condition and, so far as the situations where H_o is true are concerned, any one of them will be equally good as a critical set.

Next we must consider the case where H_o is wrong. Here we should like the critical set to possess the property that the probability of E falling within it should be as great as possible. This is a little too vague a statement, but we see at once that we are brought to consider probabilities of E

---

* J. Neyman: "Sur la vérification des hypothesis statistique composées," Bull. Soc. Math. de France, t. 63, 1935.

falling within w corresponding to cases where H_o is wrong. To be able to calculate these probabilities and to compare their values, we must make some assumptions concerning the various ways in which the hypothesis H_o may be wrong. In other words, to have the mathematical problem determinate, we must specify what is the set of admissible hypotheses that are contradictory to H_o.

Alternatively, we may specify not the whole set of admissible hypothesis contradictory to H_o, but only some of them, with regard to which we desire our test to be particularly sensitive. This step must be made, since otherwise the problem of choosing between all possible sets w, all of the same size P and all similar to W with respect to a_o, b_o, σ_1, and σ_2, will not be determinate.

Let us turn to the particular hypothesis H_o considered and see how we could specify the set Ω of admissible hypotheses contradictory to H_o. This, of course, could be done in various ways. First of all we could make a purely negative statement: if H_o is wrong, then the variables x_t and y_t do not satisfy equations (19). In this case the set Ω would consist of all possible hypotheses describing the distribution of the x and y, with the exception of H_o. Starting with this statement, it would perhaps be possible to find an appropriate test, but attempting to make it sensitive to all possible deviations from the hypothesis H_o, we shall cause it to be not sensitive enough for some in particular of them that may be regarded as especially important.

One of the possibilities in this respect is suggested by the work of Ragnar Frisch. In fact, he suggests that the machinery of the exchange between the shoemaker and the farmer should perhaps include an adjustment arising from the indebtedness of the parties. Of this, however, he is not quite sure and, when discussing examples referring to the equations (15), he finds that the most realistic of them is the one in which c is positive and d equal to zero. In such a case, the indebtedness of the farmer would influence the purchases of the shoemaker, but not those of his own.

Treating the problem from this point of view, we could consider more closely the set of hypotheses, say Ω_F, stating that the x_t and y_t satisfy equations of the form

$$\left.\begin{aligned} x_t &= (a + \alpha) y_{t-1} \\ y_t &= (b + \beta) x_{t-1} \end{aligned}\right\} \qquad (20)$$

similar to Eqs.(19) but differing from them in having either a or b or both not absolute constants but dependent on G_{t-1} (page 115) or, more generally, on t. In particular, we may have in mind situations where a and b in Eqs.(20) are of the form

$$\left.\begin{aligned} a &= a_o + a_1 t \\ b &= b_o + b_1 t \end{aligned}\right\} \qquad (21)$$

a_o and b_o, a_1 and b_1 being constant. Let us then denote by Ω_F the set of hypotheses ascribing to a_o, b_o, a_1, and b_1 any real values, and let us consider the tests that would be, in a sense,* most sensitive with respect to this

---

* See lecture III, pp. 47 and 48 in particular.

particular set of alternatives. I emphasize the fact that we do not necessarily believe that all possible hypotheses are those included in Ω_F, but we simply wish our test to be particularly sensitive to cases where one of the hypotheses in Ω_F happens to be true.

We may try to select tests of two hypotheses, H_1 and H_2, defined as follows: H_1 affirms that $a_1 = 0$ but does not specify what could be the values of a_0, b_0, and b_1. H_2 affirms that $b_1 = 0$, but does not specify what could be the values of a_0, b_0, and a_1. These hypotheses are of quite similar character, and it will be sufficient to consider only one of them, e.g. H_1.

In selecting a test for H_1, we may try to select the critical region w so that (i) whenever the hypothesis that $a_1 = 0$ is correct, then the probability of E falling within w will be P = 0.05 (say), and (ii) whenever a_1 deviates from zero, then the probability of E falling within w is increased to the greatest possible value. A test based on the critical region having these properties is unbiased of type B, and it will be most satisfactory for detecting cases in which $a_1 \neq 0$. (The critical region just referred to is described on page 121).

Performing the easy calculations described in my paper referred to above (p.121), it may easily be shown that the test under consideration is reduced to the following steps:

(a) Calculate the ratios $q_t = x_t/y_{t-1}$ for t = 1, 2, ..., n.

(b) Calculate the regression coefficient f of q_t on t.

(c) Calculate $$s_f^2 = \frac{\sigma_q^2(1-r^2)}{(n-2)\sigma_t^2} \qquad (22)$$

for an estimate of the variance of the regression coefficient of q on t, where σ_q^2 and σ_t^2 are the variances of the observed values of q and t respectively, and r the correlation between them--all calculated from the sample.

(d) Reject the hypothesis H_1 (that $a_1 = 0$) whenever $|f|/s_f > t'$ where t' is the particular value of t read from R. A. Fisher's tables for the chosen value of P(=0.05), and for degrees of freedom = n - 2.

Of course similar steps should be taken if it were desired to test the hypothesis H_2 that the coefficient b of Eqs.(20) does not depend on time, as is the situation if $b_1 = 0$.

The above steps were performed on the coordinates x_t and y_t that were used in the two curves in diagram II, page 118. H_2 was tested first. As was expected, the test did not detect any "evidence" that $b_1 \neq 0$ and thus that b in Eqs.(20) is not a constant. On the other hand upon testing H_1 (the hypothesis that $a_1 = 0$) it was found that

$$f = -0.0315 \qquad s_f = 0.0047$$

whereupon the application of rule (d) would lead to the rejection of the hypothesis that $a_1 = 0$ in Eqs.(21). If these calculations had been performed on some real data and not on an artificial example, then the next step in the research would have been with the economist, who would have to think of ways of

altering the hypothesis H_1 so as to bring it into better agreement with the observations. Perhaps he would think of Eqs.(15) or (16). If appropriate methods were at hand, then the hypothesis that he might construct could also be tested. A number of such steps, with an increasing volume of observational data, will, we hope, eventually give an econometric theory worthy of the name. However, this is a question for the future. At present, we should concentrate on preparing the necessary tools.

- - - - - - - - - - - - - - - - - - -

DR. LOUIS H. BEAN: One of the problems that the analysis of time series runs into is the fact that the successive data of x and y are themselves intercorrelated, that is, your successive values of x are correlated with one another and successive values of y are also. With that type of problem is it possible to apply the usual correlation technique, which rests, I assume, on the assumption that you have no dependence among your observations in a given series?

DR. NEYMAN: I think that if this question is taken in its full generality, then the answer is in the negative. But there are cases where the correct solution of the problem is reducible to the ordinary correlation analysis applied in some particular form. To illustrate this point we may use the example of the time series of diagram II, page 118, just discussed. It is obvious that any two consecutive values of x are in a sense correlated, their most probable variations being as in diagram I, page 116. However, as I have pointed out — and this may be confirmed by detailed calculation — the test of the hypothesis that $a_1 = 0$ is reduced to what is essentially the ordinary correlation analysis of q and t. But this circumstance is closely connected with the particular hypothesis H_0 we desired to test and with the set of alternative hypotheses with respect to which we desired our test to be particularly sensitive. I may confess here that in constructing this example I made some efforts to formulate both the hypothesis tested and the circumstances considered as known, so as to be able to arrive at the final result. It happened to be as I have described. But in the process of constructing the example I have considered a few other hypotheses to be tested and the corresponding sets of alternatives that suggested themselves as reasonable. I am sorry to say that in none of these prospective examples I was able to complete the calculations leading to the unbiased test of type B; but it was clear that these tests would <u>not</u> be covered by what is now called the correlation analysis.

My opinion is that in the question you ask no wholesale rule could be given and that each problem should be (i) properly stated: What is the hypothesis to be tested; what is considered as known, and what are the alternative hypotheses; (ii) next, each such problem should be considered and analyzed by itself. Later on, we shall certainly arrive at a classification of problems into various types, but this will not be done today or tomorrow.

Coordinates for plotting the

diagrams on pages 116 and 118.

Solution of the equations for diagram I

t	x	y
1	2.000	1.000
2	1.968	2.000
3	3.920	1.968
4	3.720	3.920
5	7.166	3.720
6	6.596	7.166
7	12.096	6.596
8	10.923	12.096
9	18.764	10.923
10	16.668	18.764
11	26.626	16.668
12	22.452	26.626
13	36.744	22.452
14	26.673	36.744
15	42.660	26.673
16	28.673	42.660
17	48.163	28.673
18	31.884	48.163
19	54.617	31.884
20	34.339	54.617
21	57.839	34.339
22	36.125	57.839
23	59.748	36.125
24	37.967	59.748
25	60.823	37.967
26	39.372	60.823
27	62.283	39.372
28	40.474	62.283
29	63.716	40.474
30	42.417	63.716

Solution of the equations for diagram II

t	x	y
1	1.770	1.111
2	2.086	1.906
3	3.921	2.376
4	4.529	3.043
5	5.127	3.714
6	7.172	4.783
7	8.480	7.502
8	12.243	8.633
9	13.019	12.096
10	18.241	13.592
11	20.334	16.472
12	18.498	20.354
13	31.325	16.926
14	28.706	25.561
15	41.639	25.864
16	23.614	34.810
17	28.022	25.125
18	25.753	29.395
19	40.830	29.384
20	43.136	37.645
21	43.668	41.324
22	60.002	50.306
23	53.878	62.882
24	71.308	50.484
25	56.189	70.452
26	65.098	56.077
27	49.236	52.860
28	52.173	47.759
29	36.536	51.860
30	62.284	36.792
31	51.214	61.101

STATISTICAL ESTIMATION,
Practical Problems and Various Attempts to Formulate their Mathematical Equivalents

A conference held in the auditorium of the Department of Agriculture, 8th April 1937, 10 a.m., Mr. Alexander Sturges, presiding.

I hope that this conference will be easier than the others because it will be mathematical. Some people think that mathematics is difficult, but whoever is acquainted with mathematics knows that it is easier than anything else because it deals with clearly defined ideas, rules, and hypotheses, that are not obscured by circumstances in life that are very, very complicated; certainly the most complicated thing in the world is life and the world outside of us.

But the problem that I am going to be concerned with this morning is connected with life; and in fact, any mathematical problem could be traced to some practical problem, as in geometry, engineering, surveying, physics, and so on. When I shall speak about the problems of estimation, I shall have two aspects in mind; one is the practical problem of estimation, and the other a mathematical model of it. What is the practical aspect of estimation? The statistician cannot study the whole population in which he is interested. This population may be, for instance, the population of farms in the United States. If for some reason it is inexpedient to study all the farms, the only thing we can do is to draw a random sample out of this population, and try to judge from it what are the properties of the population. This is a practical problem of estimation, and no question of probability is involved, nor any of mathematics. As a matter of fact it is obviously hopeless to try to get from the sample exact properties of the population. What we can hope for are some figures which "presumably" are not very wrong if treated as characters of the population.

"Presumably" is a term referring to our state of mind, but it is not a scientific term. In order to treat the problem mathematically, the mathematician must translate the requirements of practical statistics into his language. He must substitute something definite for the general idea concerning estimation. The first thing that he must do is to put the problem of practical estimation into mathematical form. In doing so he will have to deal with probability, because probability is the only mathematical concept that has something to do with the vague idea of "presumably." If we analyze the situation we shall find that the probability is counted not only at the end, but at the beginning of the problem of estimation. A sample from which the population is to be studied must be "properly" drawn; otherwise the theory of probability is not applicable. Also, we must understand the sense in which probability statements are to be interpreted. If we want a mathematical picture of the problem, we shall probably say that the statistician is

able to obtain some numbers x_1, x_2, and so on, these being measured values in a random sample. We must know something definite about the method of drawing the sample; we know also something about the population--perhaps very little; and we want to know something more about it.

The knowledge available frequently determines the general form of the integral probability law* of the x_i. However, owing to the fact that our knowledge about the population is not complete, our knowledge of the integral probability law could not be complete either. In frequent cases--and only these will be considered below--the elementary probability law* of the x_i exists, its form is known; and the only things about which we are doubtful are the values of several parameters entering the elementary probability law.

It will be sufficient to consider the case when there are only two unknown parameters, which we shall denote by θ_1 and θ_2. The case when the number of unknown parameters is greater is completely analogous. In order not to forget about the parameters that are unknown, we shall use for the elementary probability law the notation $p(E|\theta_1, \theta_2)$, which is to be the symbol for the elementary probability law of the sample point* E, calculated for the particular values θ_1 and θ_2 of the parameters. The problem of estimation is how to use the observed values of the x_i in order to estimate one or both of the unknown parameters.

The situation may be exemplified by the properties of an instrument of measurement that induces us to believe that the measurements $x_1, x_2, \ldots, x_n$ follow a normal law in which, however, the mean and the standard deviation are unknown.

What I have just said seems to be a common element in mathematical models of the problem of estimation as it is treated by various authors. In the following I shall give a short review of various ways of completing the model and thus of formulating the mathematical problem in its final dress.

The first attempt at solving the question of how to arrive at "presumable" values of θ_1 and θ_2 was based on a theorem of Thomas Bayes, published posthumously in 1763.** This applies to the situation when not only the x_i are random variables, but θ_1 and θ_2 are also; that is to say, it applies when we sample not only the values of the x_i but also the populations. We must imagine that we have a set of different populations, and out of them we pick one at random this time, another or perhaps the same one tomorrow, and so on. The formula of Bayes is familiar and I shall not go into it. I shall merely point out that it

---

* For definitions of probability laws, etc., see the first two lectures, page 16 in particular.

** Thomas Bayes, Phil. Trans. Royal Soc. London 53, 370-418, 1763.

applies only when the parameters themselves are random variables, and moreover, only when their probability distribution is known. When this is so, and we have also the observed values of the x_i forming a random sample from a population picked at random, we may apply the formula of Bayes either to calculate the posterior probability that θ_1 and θ_2 have values within any assigned limits, or we may calculate the values of θ_1 and θ_2 that are the most probable and consider them as estimates. The mathematical model of the practical problem of estimation connected with the Bayes' theorem is perfect by itself; however, there are serious difficulties with its application, which I shall now mention.

In most practical cases the parameters are not random variables. We don't sample them. Consider for instance, the population of persons who were living in Washington at a certain moment of the year 1935. This is perfectly constant and its properties are not subject to any random variation. On the other hand, if we sample randomly from this population and the x_i denote some character of the individuals drawn, then there will be no difficulty in considering the x_i as random variables. In fact, we are usually taking elaborate precautions in the method of sampling to make sure that the x_i do possess the properties of random variables according to some probability law. Nothing of this sort could be done with the properties of the population just mentioned and they must accordingly be considered as constants. Therefore, in this and in many other problems, the formula of Bayes is not applicable because it refers to a different situation, that in which the parameters themselves are random variables. Sometimes, of course, the parameters may actually be random variables, but the application of the formula of Bayes requires not only this, but also knowledge of the probability law of the parameters, since the probability law of the parameters comes into the formula and must be known.

Different scientists have advanced different ideas of how the difficulty could be overcome. This was in the early days of the theory of probability when there was a confusion between the two different elements that I spoke of yesterday evening, namely, the mathematical conception of probability and our psychological idea of probability.* They were confused very much, and a principle was advanced called the principle of insufficient reason. This principle says that whenever we do not know anything about the value of θ_1, we are allowed to assume that the probability of θ_1 lying within any interval is simply proportional to the length of that interval. The probability law of θ_1 would then be represented by a rectangle. In other words, if we have no sufficient reason to assume that some particular value of θ_1 is more probable than any other, then we may assume that the probability distribution of θ_1 is constant. On this principle, there will be no difficulty in using the formula of Bayes to calculate the most probable values of θ_1 and θ_2, or the probability that they fall within any given ranges. However, it is important to be clear about the range of validity of such calculations; they can be no better than the assumptions

---

* Cf. page 32.

put into the formula.

My second lecture was concerned with the empirical law of big numbers. You remember its general meaning: if we start with the correct mathematical model of a set of random experiments, such for example as sampling the population of Washington, and deal with probabilities as I have defined them, then whenever we find by calculation that the probability of some specified event is equal to α, the frequency of that event in repeated experiments will actually approach α. In other words, if we consider only probabilities of the kind I have described, the empirical law of big numbers permits the prediction of frequencies of future events, the probabilities of which we are able to calculate. It all depends upon having the proper mathematical model, or alternatively, arranging the experiments to fit the model.

When judging the principle of insufficient reason, it is important to remember that the assumed constant probability of θ_1 falling within any interval of a fixed size is not a probability of the kind related to the empirical law of big numbers. The former is a probability describing, as somebody has said, the state of our mind; and usually it has no relation whatever to the frequencies of θ_1 having this or that value. The consequence of this is that the most probable value of θ_1, calculated from the Bayes' theorem with the application of the principle of insufficient reason, may be described as the one that we are ready to believe in the strongest, but in general it will not be the one that we shall most frequently meet in practice. Our sampling technique in obtaining the x_i may be perfect, and the law of big numbers in everything concerning the x_i may work well,* and then, having obtained from Bayes' formula and the principle of insufficient reason that the probability of $1 < \theta_1 < 2$ is 0.999, we may be disappointed to find that the frequency of cases in which θ_1 does fall within these limits is negligible.

All this justifies further attempts to get something better than the principle of insufficient reason.

There were other principles advanced for overcoming the difficulty with Bayes' theorem. Of those, I shall mention two very briefly. One was advanced by Gauss** but not in a very clear way. It was developed and put in practical form by a famous Russian mathematician, Markoff, whose book is now translated and will appear soon in the Annals of

---

* That is, the correspondence between relative frequencies actually obtained in practice, and the probabilities calculated from a given mathematical model on the basis of a certain value of θ_i, may be very good (see Lecture I, page 18a, and Lecture II, page 22).

** See e.g., Gauss, Theoria combinationis observationum erroribus minimis obnoxiae, pars prior, page 49 (Göttingen, 1821); also J. F. Encke, Jahrbuch für 1934 (Berlin, 1832) pp. 284-285.

Mathematical Statistics.* Markoff was not a statistician, he was a mathematician. What I shall say will be exactly equivalent to what he says but it will have a form more familiar to statisticians. Suppose we are interested in the value of the parameter θ. At our disposal we have the values of x_1, x_2, etc., and they are random variables. You know the conception of mathematical expectation; I shall denote the mathematical expectation of any variable u by E(u). Now let us consider some function F of x_1, x_2, etc. Any estimate of θ will be a function of them. I shall say that the function F is an unbiased estimate of θ if the expectation of F is identically equal to θ, i.e. if $E(F) \equiv \theta$. That is to say, if whatever be the properties of the sampled population, and if whatever be the probability law of $x_1, x_2, \ldots, x_n$, the expectation of F is equal to θ, then this F will be called an unbiased estimate of θ, made from $x_1, x_2, \ldots, x_n$.

Now, I shall define what I shall call a best unbiased estimate. There are many functions whose expectation is identically equal to θ. Therefore, we are allowed to choose between several unbiased estimates. If we are to choose with a purpose, we must agree on what is the best quality. A good quality of F to consider is its standard error. What is the standard error of F? I shall define not the standard error, but the standard error squared, or the

$$\text{Variance of } F = E(F - \theta)^2 \qquad (1)$$

and I shall say that whenever the variance of F is diminished, the estimate F is improved. Then the best estimate, the best unbiased estimate will be that one of minimum variance.

Markoff has shown how we can calculate in various cases the best of the unbiased estimates which are linear functions of the x_i. Suppose, then, that we can find the function F for the best unbiased estimate; having gotten a sample we may substitute the values of $x_1, x_2, \ldots$ into the function F, and we may risk saying that the result we obtain is equal to θ, or perhaps not very much different from θ.

This statement has a justification provided by the theorem of Bienayme - Tchebycheff to the effect that the probability (in the sense of the theory of probability I am using) that the value of F will differ from its expectation θ by more than t times the standard error of F is smaller than t^{-2}, whatever may be $t > 1$. In various cases this probability may be shown to be much smaller than the limit t^{-2}.

---

* The editor understands from Professor Frank M. Weida, who is editing the translation of Markoff's book, that it will appear in supplements to the Annals of Mathematical Statistics, 50 to 75 pages in each of four or five of the quarterly issues, beginning perhaps in December 1937. Later the book will be assembled and printed as a single volume.

You will notice, however, that in spite of this justification, the use of the unbiased estimates is based on a principle and that it is not a solution of any properly stated mathematical problem. One might ask for example whether there are not some other functions of the x_i, different from the unbiased estimates, the values of which differ by a given amount from the estimated parameter still less frequently than those of the best unbiased estimate. But the dogmatic character of the unbiased estimates will be best seen from a comparison with other estimates based on a competitive principle, such as the principle of likelihood, which is again something artificially brought in and not directly inherent in the theory of probability.

QUESTION: The term unbiased estimate refers to the particular definition you have used. Does the term best unbiased estimate always refer to the two definitions that you gave?

DR. NEYMAN: Yes.* The best unbiased estimate is such a function of the sample that its expectation is equal to θ identically, and the variance of the function is smaller than that of any other unbiased estimate.

QUESTION: Can you use "unbiased estimate" basing it on some different definition not necessarily involving variance, or does best unbiased estimate always imply that particular definition?

DR. NEYMAN: In the past, the words "best unbiased estimate" have been defined in this way, but somebody may use them to describe a different conception. There is no difficulty about this. It is rather unfortunate that in the theory of probability and statistics we frequently use very suggestive terms. "Best unbiased estimate"--if people are not very sophisticated they will think that it is the best; but it is only called the best and somebody else may call something else the best. We must clearly distinguish between what is the best and what is called the best.

I shall describe now another principle, invented to solve the problem of estimation. So far as I know, it was invented in 1895 by the late leader of mathematical statistics, Karl Pearson;** he said that if the probability law of the sample depends on θ_1 (it might depend on other parameters too, but it doesn't matter--we are interested in estimating this one), then the optimum estimate of θ_1, which we may denote by $\hat{\theta}_1$, is the value of θ_1 for which the probability $p(E|\theta_1, \theta_2)$

---

* E. J. G. Pitman, in a recent article in the Proc. Cambridge Phil. Soc. 33, 212-222, April 1937 uses biased in a different sense; he says an estimate is biased if it is more frequently too large (small) than too small (large). Editor.

** Karl Pearson: "Regression, heredity and panmixia." Phil. Trans. Royal Soc. 187A, 253-318, 1895; p.265 in particular. The method of maximum likelihood seems to have been first used by Helmert, Astronomische Nachrichten 88, No. 2096, 1876. Editor.

of the observed values $x_1, x_2, \ldots, x_n$, is greater than for any other value of θ_1.

In other words, the numerical value of the probability $p(E|\theta_1, \theta_2)$ will depend on θ_1, i.e., $p(E|\theta_1, \theta_2)$ is a function of θ_1. Now according to the principle just described, from among the possible values of θ_1 I shall pick the one $(\hat{\theta}_1)$ that gives the greatest probability to the sample point E actually obtained. This is approximately the wording of Karl Pearson. He used the principle to obtain the now familiar formula for calculating the correlation coefficient, the sum of products divided by the square root of the product of the sums of squares. He said that this is the thing to do because if we assume that the population correlation coefficient ρ is equal to that in the sample, then the probability $p(E|\rho)$ of the sample will be greater than for any other value of ρ. However, Pearson did not insist upon this principle; he did not apply it in many other cases in which he was faced with the problem of estimation. This has been done with great emphasis by R. A. Fisher. He calls the expression $p(E|\theta)$ the likelihood of θ for the observed sample $E = x_1, x_2, \ldots, x_n$, and he considers its value the measure of our confidence in the particular value of θ used in the expression. The value of θ that maximizes the likelihood he calls either the optimum estimate or the maximum likelihood estimate and says this is the value in which we should have the greatest confidence.

Again, this is a new principle, saying that if you want to have the "best estimate," just calculate the value of θ that maximizes the likelihood. But you will have to believe that it is the best. This is an arbitrary principle similar to the principle of insufficient reason and similar to the principle advanced by Markoff on unbiased estimates. It is a principle that we may accept or reject just because we like it or are inclined to disbelieve it. However, there are no special reasons for believing that it is better than anything else, primarily because we did not formulate in advance what quality of an estimate we agree to consider to be of the greatest importance.

As a matter of fact, it has been possible to prove various properties of the maximum likelihood estimates (m.l. estimates, for short) that provide certain justification for their use. The justification that seems to be the main one is of the same character as the one that I quoted in favor of the best unbiased estimates. Its general effect is that, under certain limiting conditions, when all the observations are mutually independent and their number n indefinitely increases, then it becomes less and less probable that the m.l. estimate will differ by so much from the parameter that is being estimated.

It is interesting to consider the relationship of those two principles, that of Markoff of unbiased estimates, and the principle of likelihood. What is their relationship? Do they give the same results or do they give different ones? Sometimes the maximum likeli-

hood estimate is identical with the best unbiased estimate; then there is no competition between the two principles and everything is all right. Sometimes we are able to find only one of the two possible estimates, either the maximum likelihood estimate or the best unbiased estimate, simply because the equations that we need to solve to get both are too difficult and we are able to solve for only one. Here again there is no question of competition. But sometimes when we are able to determine both the maximum likelihood estimate and the best unbiased estimate, the two do not agree, and then we are in doubt which to use, and it may be difficult to choose between them.

I recall an example that is probably familiar, but in which perhaps the question of disagreement of the principles is not known to all. Suppose that each of the x_i follows the same normal probability law; then the elementary probability law of each of the x_i is

$$p(x|\mu,\sigma) = (\sigma\sqrt{2\pi})^{-1} \exp[-(x - \mu)^2/2\sigma^2] \qquad (2)$$

where both σ and μ are unknown. If the x_i are independent of one another, then the elementary probability law of all of them is*

$$p(E|\mu,\sigma) = (\sigma\sqrt{2\pi})^{-n} \exp[-\Sigma (x_i - \mu)/2\sigma^2] \qquad (3)$$

depending on the two parameters σ and μ.

Suppose that the x_i have been fixed by observation, and we wish to find a maximum likelihood estimate of σ. Differentiating (3) with regard to σ and equating the derivative to zero, we get

$$\sigma^2 = \frac{1}{n} \Sigma (x_i - \mu)^2 \qquad (4)$$

If, as is assumed, μ is unknown, then this equation does not provide us with an estimate of σ. The value of σ that would maximize the likelihood whatever be μ does not exist. The difficulty seems to be overcome if we use the principle of likelihood for a simultaneous estimation of both σ and μ. Thus we may look for a system of values of both σ and μ which would ascribe to the likelihood a value exceeding all others. As a matter of fact, this is the method by which the estimate of correlation was found. Differentiating the likelihood with respect to μ and equating the derivative to zero, we obtain the equation

$$\mu = \frac{1}{n} \Sigma x_i \equiv \bar{x} \qquad (5)$$

This equation simultaneous with (4) gives us the maximum likelihood estimates of both μ and σ^2, the latter being

$$\sigma^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2 \equiv s^2 \qquad (6)$$

---

* Throughout this conference the subscript i in the summations will be understood to run from 1 to n, i.e., over the whole sample.

But as is well known, the best unbiased estimate of σ^2 is (say)

$$s'^2 = \sum (x_i - \bar{x})^2/(n-1) \qquad (7)$$

Thus the two estimates are not equal; maximum likelihood gives s^2, while the "best unbiased" estimate is $s'^2 = ns^2/(n-1)$.

The difference is small, but it exists; and which of the two estimates is a question that has no meaning. It is remarkable that some people who say they believe strongly in maximum likelihood estimates, in actual practice use the best unbiased estimates of σ^2. It is deplorable that it is not understood that the question whether maximum likelihood or the best unbiased estimate should be used is one of taste only. If this were clearly understood then there would be no room for unnecessary polemics. The people would see that the choice between s^2 and s'^2 is more or less that between the scents of Coty and Houbigant, or between wines from Bordeaux and Bourgogne, or between "Scotch" and "Irish."

DR. DEMING: There is nothing unique about a best unbiased estimate. An unbiased estimate of σ^2 gives a <u>biased</u> estimate of σ or of σ^3, or any function of σ other than σ^2. Likewise, an unbiased estimate of σ gives a biased estimate of σ^2, etc.

DR. NEYMAN: You are perfectly right and this is just one of the arguments advanced in favor of the maximum likelihood estimates. These have the property that if T is a maximum likelihood estimate of θ, then T^2 will be one of θ^2, etc. There are many such properties of the two kinds of estimates which are sometimes considered as "proving" that one or another is better. I doubt, however, whether they are very persuasive. With respect to what you have mentioned one could ask for example, <u>why</u> should we require that the estimate of σ should be the square root of that of σ^2 ? If our purpose is to estimate σ^2 and we like unbiased estimates, then we should use s'^2. On the other hand, if we are interested in σ, we should use

$$s'' = s\sqrt{(\tfrac{1}{2}n)}\ \Gamma(\tfrac{1}{2}\overline{n-1})./\ \Gamma(\tfrac{1}{2}n) = s\sqrt{(1/2\pi)}\ B(\tfrac{1}{2}\overline{n-1}, \tfrac{1}{2}) \qquad (8)$$

which will be an unbiased estimate of σ.

It may be mentioned also that while s^2 is a maximum likelihood estimate of σ^2 <u>only</u> if the variables considered are normally distributed, s'^2 has the property of being an unbiased estimate of σ^2, <u>whatever</u> be the distribution of the x_i, just so $\sigma^2 < \infty$. This is an argument in favor of s'^2. As I have said, there are many important properties that are frequently quoted to support one or the other of the two principles --so many that there is hardly time enough to enumerate them all.

MR. FRIEDMAN: It is true, is it not, that if you take the distribution of the sum of squares ns^2, and get the maximum likelihood estimate

of σ from that, you will get $n - 1$ in the denominator?*

DR. NEYMAN: Yes, it is true; but if you imply that the maximum likelihood estimate of σ is what I have denoted by s', and not s, because of the circumstance you mention, then I should disagree. Or rather I would say that the definition of a m.l. estimate, as you seem to interpret it, is not sufficiently categoric to apply to one unique statistic at a time. Taking the distribution of ns^2 and maximizing it to get the m.l. estimate of σ implies that this must be a function of the sum of squares $\Sigma (x_i - \bar{x})^2$. If you do this, I would ask why you do not start with the distribution of the mean deviation, say $M = (1/n)\Sigma|x_i - \bar{x}|$, or of that of the range $R = x_n - x_1$, x_1 and x_n being the smallest and the largest of the n observations. In each case you will be able to find a maximum likelihood estimate of σ, but one of them will be a function of M, and the other a function of R. They would be different, and both would differ from s', and the question would arise which to choose. The choice would require some new principle in addition to that of maximizing the likelihood.

My impression is that the originator (footnote page 132) of m.l. estimates had in mind taking the original elementary probability law of the x_i, in a form like Eq.(3), and finding the values of the parameters which, for a given sample point E, would maximize it.

MR. FRIEDMAN: If you more or less make "double" application of maximum likelihood whenever you don't come out with the best unbiased estimate, later sometimes you do. First, you get s^2 (see Eq. 6); then if you apply maximum likelihood to the distribution of s^2, you will get $ns^2/(n - 1)$, an unbiased estimate* of σ^2. If you now take this estimate and get its distribution and apply maximum likelihood again, you come back to the same unbiased estimate of σ^2.

DR. NEYMAN: My point is that if you start with your original distribution of the x_i, and if you maximize the probability so as to obtain the original maximum likelihood estimate, this estimate has not necessarily the properties of the unbiased estimate. By your device, you may obtain something that will have the property of being an unbiased estimate, but then it will not have the property of maximizing the likelihood.

In order to illustrate the dogmatic character of the two principles of unbiased and the m.l. estimates, I will give one more example.

Consider a case where it is known that all the x_i that may be given by observation are mutually independent, and that each of them follows the same probability law

---

* This was Helmert's way of arriving at $s'^2 = ns^2/(n - 1)$ for an estimate of σ^2 in 1876; see reference cited on page 132. Editor.

$$p(x|\theta) = \theta^{-1} \quad \text{for } 0 < x < \theta$$
$$p(x|\theta) = 0 \quad \text{for any other value of } x. \tag{9}$$

This is what is called a rectangular distribution with an unknown range θ, starting with zero. It is desired to estimate θ.

The maximum likelihood estimate of θ is the greatest of the x_i observed in a sample, and I shall denote this estimate by g. The elementary probability law of g is easily found to be

$$p(g|\theta) = \theta^{-n} n g^{n-1} \quad \text{for } 0 < g < \theta \tag{10}$$

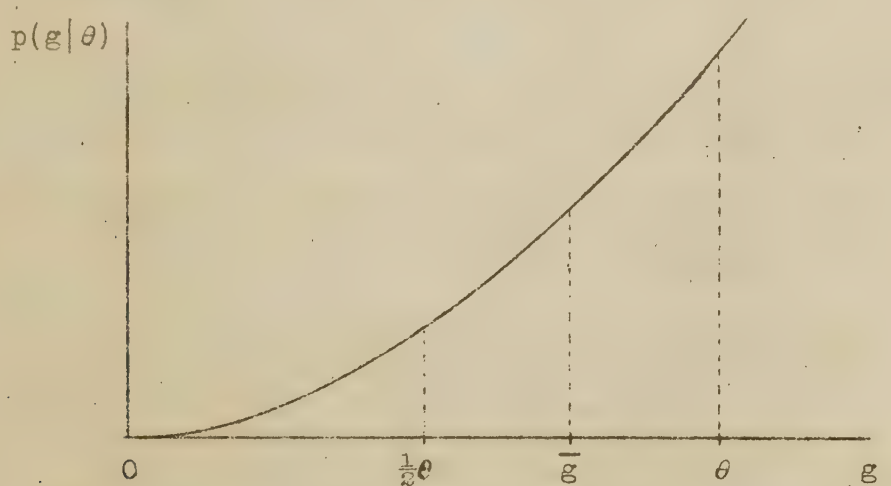
and zero elsewhere, where n denotes the size of the sample, as before. The mathematical expectation of g is

$$E(g) = n\theta(n + 1)^{-1} = \bar{g} \tag{11}$$

so that g is not an unbiased estimate of θ. An unbiased estimate, say g_1, will be provided by

$$g_1 = (n + 1)\, g\, n^{-1} \tag{12}$$

Which of the two should we use? Consider more closely the particular case where n = 3, so that the elementary probability law of g is represented by a parabola. To take g as an estimate of θ is to assume that the observed g is exactly equal to θ.



Judging from the graph this, of course, could be roughly described as "the most frequent" value of g, but it is most certain that in any practical case g will be smaller than θ. It is certain also that g cannot exceed θ, so that using g as an estimate, we are bound to make errors always of the same sign. This may induce us to use g_1 as an

estimate rather than g. Doing so we shall make errors sometimes positive and sometimes negative, and the average of these errors would tend to become zero. But, on the other hand, to take g_1 as an estimate is to presume that the value of g we observe is equal to $\overline{g} = \frac{3}{4}\theta$ (see Eq.11). If we look at the diagram representing $p(g|\theta)$, we shall see that it is so much more probable that g will fall between $\overline{g}$ and $\overline{g} + \frac{1}{4}\theta$ than between $\overline{g}$ and $\overline{g} - \frac{1}{4}\theta$. This may be considered strongly against g_1.

Shall we use as an estimate something half way between g and g_1? But here still another consideration comes in. The probability of g exceeding $\overline{g}$ is $1 - (\frac{3}{4})^3 = 0.578$, and therefore the probability of its falling short of $\overline{g}$ is 0.422. Wouldn't it be better to assume that the value of g we observe is (say) $\overline{\overline{g}}$, so defined that the probability of its being exceeded by chance is exactly equal to 0.5? It is easily seen that generally

$$\overline{\overline{g}} = \theta \sqrt[n]{\tfrac{1}{2}} \tag{13}$$

and that, therefore, the corresponding estimate of θ would be (say)

$$g_2 = g \sqrt[n]{2} \tag{14}$$

It is seen that we are involved in a dispute without a backbone. Any suggestion to use g or g_1 or g_2 is a dogma, and anyone may choose the one he likes and call it "best." Until we have found a generally acceptable form of the problem of estimation, it is useless to insist that one or the other of the suggested estimates is the best.

Attempting to reach an acceptable solution of the problem of estimation, we must bear in mind all the circumstances of the problem and its aims. The relevant points seem to be as follows.

1° Any attempt to estimate a parameter θ implies the desire (i) to make a statement concerning the value of θ and (ii) to avoid errors in this statement.

2° Any statement concerning θ will have to depend on the values of some random variables provided by observation. Therefore, if the statement is made according to some rule, it will be a function of the random variables and in consequence it will have the property of randomness; it will be subject to the law of great numbers. This means that, knowing the probability law of the x_i, we may attempt to calculate probabilities of our statement having this or that property and in particular of its being correct. If it is found that the probability of a statement concerning θ is 0.99, when made according to a specified rule, then we shall know that, in the long run, the rule will actually lead to correct statements in about 99 percent of all cases applied.

3° From this point of view, it is hopeless to look for a

solution of the problem of estimation in the form

$$\theta = \text{some specified function of the } x_i = T(E), \text{ say.}$$

In fact, whatever the function T(E), if θ can possess any value out of a finite or infinite interval, the probability of T(E) being exactly equal to θ must be zero. It follows that, in most cases, the problem of a unique estimate treated from the point of view of 1° and 2° has no solution. But, as a matter of fact, the practical statistician must be aware of this circumstance and moreover, what he is actually doing suggests that he has already given up the idea of a unique estimate. If you look through any number of recent statistical publications, you will find that the results of estimating means, correlations, etc. are invariably given in the same form: $T \pm S_T$, where T is an estimate deduced from this or that principle and S_T the estimate of its standard error. This manner of writing and also the comments on the results suggest that the practical statistician has in mind indicating an interval extending from T minus some more or less vaguely specified multiple k_1 of S_T to T plus some other multiple k_2 of S_T, in which "presumably" the true value of the estimated parameter is contained.

This last circumstance is the main part of my point 3°, which we have to remember when formulating the problem of estimation: the form of statement that it is desired to make concerning the value of θ is

$$\underline{\theta}(E) \leq \theta \leq \overline{\theta}(E) \tag{15}$$

where $\underline{\theta}(E)$ and $\overline{\theta}(E)$ are some functions of the x_i. The familiar $T - k_1 S_T$ and $T + k_2 S_T$ are only traditional forms of these functions and it is difficult to say in advance whether and when they are satisfactory.

4° If it were possible to define more than one pair of functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ both having the property that for purposes of estimating θ in the form (15) we shall be correct in a fixed and sufficiently large percentage of cases, then we could choose the one that conforms with our view on the accuracy of estimation. Frequently, but not always, it will be the pair giving generally the narrowest intervals, i.e., the pair that makes the interval

$$\overline{\theta}(E) - \underline{\theta}(E) \tag{16}$$

as narrow as possible.

It will be noticed that the four above points differ essentially from the principles advanced as solutions, or rather as substitutes for

a solution, of the problem of estimation. In fact none of the four points is dogmatic.

The points 2° and 3° simply describe the situation and they do not contain any "you should do this or that." The other two points do contain something of that sort; namely, point 1° contains, "you should try to make erroneous statements as rarely as possible," and point 4°, "you should try to make your statements concerning θ as precise as possible" in the sense that you would prefer

$$1 \leq \theta \leq 2 \tag{17}$$

rather than

$$1.5 \leq \theta \leq 5 \tag{18}$$

but these "you should" are not dogmas. Whoever takes the trouble to make some observations and to work them out mathematically, must have these two "you should" in mind. Otherwise, he would probably offer an estimate of any parameter simply by opening a book of logarithms and reading the first figure that his eye would fall on, or the like.

The first three of the above points lead to a mathematical problem as follows:

Knowing the probability law $p(E|\theta_1,\theta_2)$ of the x_i, the problem of estimating θ_2 is to determine two functions of the x_i, namely $\underline{\theta}(E)$ and $\overline{\theta}(E)$, satisfying the condition

$$\underline{\theta}(E) \leq \overline{\theta}(E) \tag{19}$$

and such that, if $\theta_1°$ is the true value of the parameter θ_1, then the probability of $\underline{\theta}(E)$ satisfying the inequality

$$\underline{\theta}(E) \leq \theta_1° \tag{20}$$

and at the same time of $\overline{\theta}(E)$ satisfying the inequality

$$\theta_1° \leq \overline{\theta}(E) \tag{21}$$

is identically equal to a number α, close to unity, and chosen in advance. (This α is different from the α on pp. 49-88)

Point 4° indicates how to choose between the solutions of this problem if there is more than one. It must, however, be made more precise.

This is the mathematical problem of statistical estimation as I understand it, and in my next conference I will give you some indications toward its solution.

MR. PAGE: I have a very simple question; I didn't understand

clearly the maximum likelihood method of estimating σ^2. Could you take the square root of the estimate of σ^2? Does that give you the maximum likelihood of σ itself?

DR. NEYMAN: Yes. Say T = max. likelihood estimate of θ.
Then T^2 = " " " " θ^2
$\sqrt[n]{T}$ = " " " " $\sqrt[n]{\theta}$

If I consider as my parameter, not θ but any function of θ, then the maximum likelihood estimate will be the same function of T. This does not apply to the unbiased estimate.

MR. WALLIS: Doesn't Fisher claim that maximum likelihood solutions will always be minimum-variance solutions also? I thought that Fisher claimed that he would get the "best" estimate by the method of maximum likelihood.

DR. NEYMAN: I am aware of these claims.* However, the proofs advanced by Professor Fisher to support them were not considered satisfactory by many mathematicians and recently several interesting papers have appeared on the subject. As a result, many of Fisher's statements, partly in a modified form and under certain limiting conditions, proved to be correct. I do not remember whether the particular claim you mention was found correct or wrong, but I will quote here papers by Hotelling, Doob, Dugué, and Pitman, where you are likely to find the answer.**

But my point is that the question whether the variance of the m.l. estimate is minimum or not is not relevant from the point of view of the goodness of the estimate itself. In the above example, the variance of g is smaller than that of g_1, but does this circumstance prove the absolute superiority of g over g_1?

---

* R. A. Fisher, Messenger of Mathematics 41, 155-160, 1912. Phil. Trans. Royal Soc. 222A, 309-368, 1922; Proc. Cambridge Phil. Soc. 22, 700-725, 1925; Proc. Royal Soc. 144A, 285-307, 1934; J. Royal Stat. Soc. 98, 39-82, 1935.

** Harold Hotelling, Trans. Amer. Math. Soc. 32, 847-859, 1930.
J. L. Doob, ibid. 36, 759-775, 1934; 39, 410-421, 1936; also Annals of Math. Stat. 6, 160-169, 1935.
Daniel Dugué, Compte rendus 202, 193-195, 452-454, 1732-1734, 1936.
E. J. G. Pitman, Proc. Cambridge Phil. Soc. 32, 567-579, 1936; ibid. 33, 212-222, 1937.

Having reread the above draft of the conference, I find that it may suggest the idea that in my opinion all the previous work on the theory of estimation is more or less useless. I want to emphasize that this suggestion would be entirely wrong. All the attempts to treat the problem in this or that way were the necessary steps marking an advance and permitting further steps ahead. And probably still for a long time all of us will calculate sometimes a best linear estimate and use the Markoff theorem, sometimes the m.l. estimate and sometimes, whenever circumstances permit, the Bayes' formula.

Apart from this I must point out a remarkable circumstance that we may frequently notice in many branches of science. This is that the uncontrollable intuition of the practical worker suggests to him the proper solution of his problem while he is entirely helpless in giving reasons why he is proceeding in this or in that way. If he is pressed for these reasons he frequently produces a principle that has the appearance of proving something, but which actually proves nothing. To illustrate my point, I will mention the familiar history of vaccination invented, as they say, by country women long before anything like modern serology had been started.

The rôle of a rigorous scientific theory is frequently very modest and is reduced to explaining to the practical man--and this sometimes with certain difficulty--how good is what he knew himself to be good long ago.

In particular, the theory of estimation, of which I will speak in more detail in the next conference, shows that many of the familiar formulas like $T \pm S_T$ are the best that could be formed: proceeding in this way, we get statements concerning the estimated parameters in their most exact form, and also we attain the relatively greatest frequency of correct statements.

This is an illustration of a statement by Laplace which I like very much: "La théorie des probabilités n'est au fond que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte." After this quotation one might ask perhaps whether theory is of any use at all to the practical man. I think it is. He is occasionally in doubt and then a theory is useful. Sometimes also his ineffable instinct is actually misleading.

AN OUTLINE OF THE THEORY OF CONFIDENCE INTERVALS

A conference with Dr. Neyman in the auditorium of the Department of Agriculture, 9th April 1937, 10 a.m., Dr. Frederick V. Waugh, presiding.

This morning I shall start with the problem of estimation as we formulated it at the close of yesterday's conference (page 140), this being the form that I consider the proper approach. It will be a mathematical problem.*

Consider n random variables $x_1, x_2, \ldots, x_n$, dependent or independent, the values of which we can observe. Those n observations will determine a point E (see page 16). Suppose that the probability law of the sample E (i.e. of $x_1, x_2, \ldots, x_n$), though known, is written in terms of some two parameters θ_1 and θ_2 which are not known. There may be other parameters also, but for simplicity I shall consider that there are only two; when there are more, the situation is similar. I want to use the sample to make an estimate of one of the parameters, say θ_1.

When I say that I want to estimate θ_1, I mean that I want to have a way of calculating two functions of E, the sample point $x_1, x_2, \ldots, x_n$ - one function to be denoted by $\underline{\theta}(E)$ and the other by $\overline{\theta}(E)$. Those will be called the lower and the upper estimates** of θ_1. I shall now describe the properties that I want the functions $\underline{\theta}$ and $\overline{\theta}$ to possess. You notice first of all that, being functions of the sample $x_1, x_2, \ldots, x_n$, they are both random variables, and both will vary from one sample to another as the sample point $x_1, x_2, \ldots, x_n$ varies. Since they are random variables, I may consider the probabilities of $\underline{\theta}$ and $\overline{\theta}$ lying within or without any specified ranges.

Let us denote by θ_1° the value of the parameter θ_1 which in my particular problem happens to be true. I don't know what this value is but I denote it by θ_1°. Now one of the properties that I want the two thetas to possess is this: I want the probability‡

$$P\{\underline{\theta}(E) < \theta_1^\circ < \overline{\theta}(E) \,|\, \theta_1^\circ, \theta_2\} = \alpha \text{ (e.g. 0.95 or 0.99)} \qquad (1)$$

---

* The reader will realize that the problem of estimation was touched upon earlier; see page 29.

** These are conveniently read "theta lower" and "theta upper". $\underline{\theta}(E)$ and $\overline{\theta}(E)$ will occasionally be abbreviated $\underline{\theta}$ and $\overline{\theta}$, the letter E being omitted for brevity. But they are nevertheless functions of E, i.e. of the sample.

‡ To be read "the probability that when θ_1° and θ_2 are the true values of the parameters, $\underline{\theta}(E)$ is less than θ_1° and $\overline{\theta}(E)$ is greater than θ_1°, is equal to α."

In words, the interval $\underline{\theta}$, $\overline{\theta}$ is to overlap the true value of θ_1° with frequency α, which I choose myself, as for instance, 0.99, or something similar.

If I succeed in finding the functions $\underline{\theta}$ and $\overline{\theta}$ satisfying (1), then I shall call the number α the confidence coefficient, and the interval extending from $\underline{\theta}(E)$ to $\overline{\theta}(E)$ the confidence interval corresponding to the sample point **E**.

Now, I shall emphasize something concerning Eq.(1) which I purposely deferred--I did not mention it clearly in order to have the opportunity of emphasis. You remember I said that θ_1° denotes the value of θ_1 that happens to be true, but I do not know what it is. (If I knew what the true value of θ_1 is, then there would be no question of estimating). On the left-hand side of equality (1) I have only a symbol for θ_1° and I can't put there any number instead of θ_1° because I don't know what this number is. Therefore, this equation really should be considered not merely as an equation but as an identity; it is to hold for all values of θ_1, as any one of them may happen to be the true one. But this is not all. You remember that we assumed that the elementary probability law $p(E|\theta_1, \theta_2)$ depends on two parameters θ_1 and θ_2, both unknown. It follows that I desire the functions $\underline{\theta}$ and $\overline{\theta}$ to satisfy (1) also identically for all possible values of θ_2.

At first sight, this problem seems to be not an easy one, and not of a usual kind. What we have in all sorts of books on probability are formulas giving the probability that certain functions of the sample, such as $\overline{x}$, or $\overline{x}/s$, or $\overline{x}/\sigma$, will fall below a certain number, or exceed a given number, when the properties of the probability law are all specified. In the present problem, the situation is more complicated; we require a probability, calculated from a probability law depending upon θ_1 and θ_2, to have a specified value, whatever the values of θ_1 and θ_2. The ultimate solution, however, is easily obtained.

However, before going into the details of the solution I want to make its purpose entirely clear. Suppose for a moment that we have succeeded in calculating the functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ satisfying (1) identically, and let us see how we could use them to produce a solution of the practical problem of estimating θ_1. The practical statistician is able to observe the x_i, and he wishes to know how these should be used for making some statement concerning the value of θ_1.

We may advise him to perform the following three steps which, together, are equivalent to a single random experiment:

(i) to observe the values of the x_i, called E.

(ii) to calculate the corresponding values of the functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ and

(iii) to state that $\underline{\theta}(E) \leq \theta_1 \leq \overline{\theta}(E)$.

You will notice that in this statement he may be correct or he may be wrong. But, owing to the properties of the functions $\underline{\theta}$ and $\overline{\theta}$ as expressed by Eq.(1), the probability of his being correct will be equal to α (e.g. 0.99). It follows that if the experiment is so arranged that the x_i do follow the elementary probability law that served for constructing the functions $\underline{\theta}$ and $\overline{\theta}$, then the empirical law of big numbers will guarantee that the practical statistician following the above advice will be correct in his statements concerning the value of θ_1 in 99 percent of all cases.

The situation may be compared with that of a game of chance in which the probability of winning has a fixed value. In the case of roulette, for example, the gambler is allowed to bet in various ways, but whatever he may choose, the probability of the bank winning the game is fixed in advance, to the greater or smaller advantage of the bank. The uncontrollable choice of the gambler of how to bet corresponds to the possibility of θ_1 having this or that value. The choice of the rule for calculating the functions $\underline{\theta}$ and $\overline{\theta}$ corresponds to fixing the rules of the game, assuring that the probability of the bank winning (= the probability of the statistician making a correct statement on the value of θ_1) is fixed in advance and sufficiently large. The choice of the functions $\underline{\theta}$ and $\overline{\theta}$ is made according to the particular probability law $p(E|\theta_1, \theta_2)$ which the x_i are assured to follow, and correspondingly the rules of the game of roulette are fixed under the assumption that the ball stops equally frequently at each of the sectors. The actual frequencies of successes in both of the "games" depend essentially on whether these assumptions are, for each one, sufficiently well satisfied.

MR. FRIEDMAN: Your statement of probability that he will be correct in 99 percent of the cases is also equivalent to the statement, is it not, that the probability is 99 out of 100 that $\theta_1°$ lies between the limits as given by $\underline{\theta}$ and $\overline{\theta}$?

DR. NEYMAN: No.* This is just the point I tried to emphasize in my first two lectures both in theoretical discussions and in examples. $\theta_1°$ is not a random variable. It is an unknown constant. In consequence, if you consider the probability of $\theta_1°$ falling within any limits, this may be either zero or unity, according to whether the actual value of θ_1 happens to be outside of these limits or within. The position is exactly as in my example on page 5 with the 1000th figure in the expansion of π = 3.14159..., which I denoted by x_{1000}. You may remember that if any calculations led to the conclusion that the probability $P\{1 < x_{1000} < 5\}$ has a value different from both zero and unity, then these calculations are either wrong or else they must refer to some other theory of probability different from the one I am using. The

---

* See, however, the editor's footnote on page 146. The point is that we must not speak of the probability of θ_1 lying within fixed limits, nor limits that are not random variables. Editor.

connection with the empirical law of big numbers is to my mind a sufficient reason to deal essentially with the particular theory of probability that I have chosen. In consequence, in everything I said, and in what I am going to say, there is no room for a probability of θ_1 having this or that value or fulfilling any given inequality, wherein this probability is other than 0 or 1.

Referring to the "picturesque" way of speaking, according to the conventions established in my first two lectures, we may say that neither θ_1° nor x_{1000} is a random variable because their values do not depend upon any random experiment. However, we could consider random variables, say θ_1 and x, having certain connections with θ_1° and x_{1000}; namely, we could define a method of picking up at random any one of the ten digits 0, 1, 2, ..., 9 for a particular position (as e.g. the 1000th decimal) in the expansion of π. Any digit picked up could be equal to the 1000th decimal, x_{1000}. Moreover, with regard to any one of the digits picked up, say x, we could perhaps assert that $P\{1 \leq x \leq 5\} = \frac{1}{2}$ or the like, depending on the behavior of the experiment. For instance, it would be possible, by exercising sufficient care, to arrange the experiment so that the frequency of the numbers 1, 2, 3, 4, and 5, approaches closely $\frac{1}{2}$ the total number of draws (see page 18a).

Similarly, we could consider a method of picking up at random probability laws $p(E|\theta_1, \theta_2)$, differing among themselves by the values of the parameters θ_1 and θ_2, one of these values being $\theta_1 = \theta_1^\circ$. The specification of this experiment method would be equivalent to the definition of the probability law of θ_1 and θ_2, from which we could calculate probabilities of their falling within any specified limits. In this way we should most certainly fall back on the calculations connected with Bayes' theorem (page 128). As I said yesterday, this theory is faultless by itself, but its applications are rare, because it is unusual for the probability laws of θ_1 and θ_2 to be known. In fact, we may assume, to a certain extent, the random character of the x_i, since the method of obtaining them is frequently under our control; but the variation of the values of the parameters θ_1 and θ_2 is usually beyond our control. Those are the reasons why the probability of $\underline{\theta}(E)$ falling below θ_1° and $\overline{\theta}(E)$ falling above θ_1°, $\underline{\theta}(E)$ and $\overline{\theta}(E)$ being random variables, is not equivalent to the probabilities of θ_1° falling within any assigned (fixed) limits.*

---

* I think Dr. Neyman would agree that one could speak of the probability of $i < x_{1000}$, or of $x_{1000} < j$, or both, if i and j are to be chosen at random according to some specified scheme of chance. The value of $P\{i < x_{1000} < j\}$ could range between 0 and 1, and would depend on the rules (i.e. the experiment) by which i and j are to be drawn. Editor.

Now we shall go on to see how in general we could obtain functions such as $\underline{\theta}(E)$ and $\overline{\theta}(E)$ satisfying (1). If I want to obtain the general way of determining the lower and upper estimates, $\underline{\theta}$ and $\overline{\theta}$, the thing to do is to assume that I already have solved the problem and see what way of constructing those functions would be followed.

I think I talked to you about the sample space.* I shall repeat it now. We denote by E the system of values $x_1, x_2, \ldots, x_n$, which we can observe. E was described as a sample point in a space of n dimensions.** This sample space will be denoted by capital W. I must say that W is not necessarily the whole space but it is the set of different possible positions of sample points. If each x_i is normally distributed, or distributed on any continuous curve, the possible sample points form an n dimensional continuum. Or, it may be that the possible positions of E are limited to certain discrete points of the space, as will happen if each x_i can take certain discrete values, but nothing between. Anyway, W will denote the set of all possible positions of E.

Now I shall consider something which I shall denote by capital G, the general space. The points in this space will be denoted by g. They have n+1 coordinates--the possible values of $x_1, x_2, \ldots, x_n$, and also the value of θ_1. Now if I fix the value of θ_1 momentarily so that

$$\theta_1 = \theta_1' \qquad (2)$$

I shall obtain a section of the general space, a plane whose equation is $\theta_1 = \theta_1'$. Any point g having coordinates $x_1, x_2, \ldots, x_n$, and θ_1', will lie on this plane no matter what be the values of $x_1, \ldots, x_n$. It is seen that there is a one to one ocrrespondence between the sample points E in W and the points g on any plane fixed in G by Eq.(2).

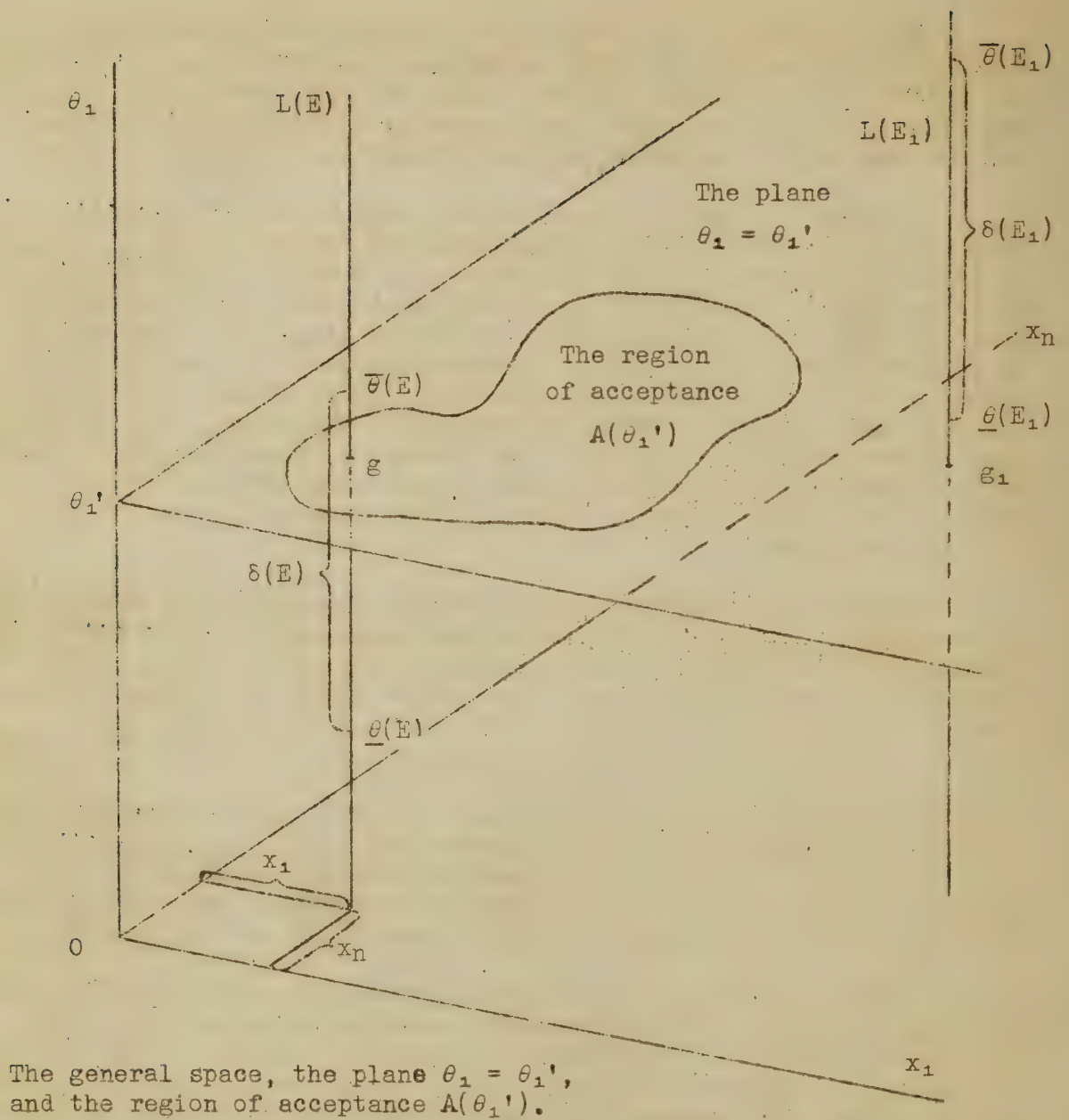
However, if we take into consideration the sample space W on the one hand, and the whole general space G on the other, we shall find that to any point E in the former corresponds a straight line, say L(E), in the latter, parallel to the θ_1 axis. Along such a line L(E) in the general space, the parameter θ_1 takes on all possible values proper to its nature, while the sample $x_1, x_2, \ldots, x_n$ remains constant. The situation is illustrated in the diagram on page 148.

The lines L(E), each corresponding to a particular sample E, are necessary for the geometrical interpretation of the functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$, the method of calculation for which we have assumed to be known. Take any sample E, find the corresponding line L(E) and, having calculated the values $\underline{\theta}(E)$ and $\overline{\theta}(E)$, plot them on the line L(E). The interval between the two points plotted will be denoted by $\delta(E)$ and called the confidence interval corresponding to the particular sample E.

---

* The reader may refer back to page 16.

** Again it is convenient to recall that a "point" is a set of numbers.



The general space, the plane $\theta_1 = \theta_1'$, and the region of acceptance $A(\theta_1')$.

A set of n observations $E = x_1, x_2, \ldots, x_n$ determines the line L(E). This line pierces the plane $\theta_1 = \theta_1'$ at the point g. Note that point g lies within the region of acceptance $A(\theta_1')$ and that the confidence interval $\underline{\theta}(E)$, $\overline{\theta}(E)$ covers θ_1'. The line $L(E_1)$ arises from some other sample. It pierces the plane $\theta_1 = \theta_1'$ at the point g_1. Note that point g_1 lies <u>outside</u> the region $A(\theta_1')$ and that the confidence interval $\underline{\theta}(E_1)$, $\overline{\theta}(E_1)$ does <u>not</u> cover θ_1'.

This can be done for all samples, and for each one we shall have in the general space a line L(E), and on that line a confidence interval $\delta(E)$ extending from the value of $\underline{\theta}(E)$ to the value of $\overline{\theta}(E)$. Every new sample gives a new line L(E) and a new pair of values for $\underline{\theta}(E)$ and $\overline{\theta}(E)$, and hence a new confidence interval.

Now I shall define something that I shall call the region of acceptance corresponding to a given value of θ_1. Again, fix the value of θ_1 at, say, θ_1'; and consider the plane in the general space corresponding to this constant value of θ_1--in other words, the plane defined by Eq.(2). If I take any sample $E = x_1, x_2, \ldots, x_n$, and the corresponding confidence interval $\delta(E)$ lying between $\underline{\theta}(E)$ and $\overline{\theta}(E)$ on the line L(E), I shall find that the plane $\theta_1 = \theta_1'$ does one of two things --it cuts the line L(E) either <u>interior</u> to the confidence interval $\delta(E)$, or else <u>exterior</u> to it. If the former is the case, it can be said that for this particular sample, $\underline{\theta}(E)$ and $\overline{\theta}(E)$ satisfy the inequalities

$$\underline{\theta}(E) \leq \theta_1' \leq \overline{\theta}(E) \qquad (3)$$

In other words, the point at which the plane $\theta_1 = \theta_1'$ cuts the line L(E), for this particular sample, lies <u>between</u> $\underline{\theta}(E)$ and $\overline{\theta}(E)$. For some other sample the situation may be otherwise, as the figure on page 148 illustrates.

Consider now all possible samples, and hence all possible lines such as L(E), each with its confidence interval $\delta(E)$ properly calculated. The plane $\theta_1 = \theta_1'$ will cut some of these lines at points interior to the confidence intervals, and the set of all such points will be denoted by $A(\theta_1')$, to be called the <u>region of acceptance</u> corresponding to the value θ_1' of θ_1. To be precise, the set of points $A(\theta_1')$ does not necessarily form a continuous region in the plane $\theta_1 = \theta_1'$, bounded by a closed curve, but in most practical problems it is so, which justifies my terminology. Generally speaking, the region of acceptance $A(\theta_1')$ corresponding to θ_1' is determined by the equation $\theta_1 = \theta_1'$ and the inequalities (3).

I shall have to introduce some more new conceptions and notations. Take into consideration some particular value θ_1' of θ_1 - not necessarily the true one - and some particular sample point E. If it happens that $\underline{\theta}(E)$ and $\overline{\theta}(E)$ calculated for this sample lies on opposite sides of θ_1' so that inequality (3) is satisfied, then I shall say that the confidence interval $\delta(E)$ corresponding to E <u>covers</u> the value θ_1'. This will be denoted by the symbol

$$\delta(E)\ C\theta_1' \qquad (4)$$

Again, if the coordinates of the sample E determine a point on the plane $\theta_1 = \theta_1'$ that belongs to the region of acceptance $A(\theta_1')$, then I shall say that the sample point E in the sample space W (of n dimensions) <u>falls within the region of acceptance $A(\theta_1')$</u> corresponding

to the value θ_1' of θ_1. This will be denoted by

$$E \in A(\theta_1') \tag{5}$$

With this notation, standing for "E is an element of $A(\theta_1')$," you are, of course familiar from my previous lectures and conferences.

Now you will notice that the two events, denoted by (4) and (5) respectively, are identical: whenever the confidence interval $\delta(E)$ covers any particular value θ_1' of θ_1, then the corresponding sample point E must "fall within the region of acceptance $A(\theta_1')$," and inversely. In fact, (4) and (5) are two different ways of describing the same thing. Owing to the circumstance that there exists a one to one correspondence between the sample points in W and the points g on any fixed plane $\theta_1 = \theta_1'$, any region of acceptance $A(\theta_1)$ on such a plane will correspond to some particular region in W, which may also be denoted by $A(\theta_1')$. It follows that the event (4) is equivalent to "E falls within the region $A(\theta_1')$ in the sample space W."

Until the present time we have not considered probabilities and have not made any assumption of what the true value of θ_1 actually is. So now we shall consider some probabilities.

If you notice that the two events (4) and (5) happen or fail together, you will agree that the probabilities of the two events are the same, whatever be the true values of θ_1 and θ_2. I may therefore write

$$P\{\delta(E) \subset \theta_1' \,|\, \theta_1, \theta_2\} \equiv P\{E \in A(\theta_1') \,|\, \theta_1, \theta_2\} \tag{6}$$

This equation is to be read "the probability that $\delta(E)$ covers θ_1', is identical with the probability that E is an element of $A(\theta_1')$, whatever be the true values of θ_1 and θ_2."

You will recall that we started with the assumption (Eq.1) that the two functions $\underline{\theta}$ and $\overline{\theta}$ are calculated by a rule that makes the probability of $\underline{\theta} < \theta_1^\circ < \overline{\theta}$ equal to the confidence coefficient α, whatever be the true value θ_1° of θ_1 and whatever be the value of θ_2. In other words we started with the assumption that

$$P\{\underline{\theta}(E) < \theta_1^\circ < \theta(E) \,|\, \theta_1^\circ, \theta_2\} \equiv \alpha \tag{1}$$

So now, with what we have just seen in (6), we may say that

$$P\{\underline{\theta}(E) < \theta_1^\circ < \overline{\theta}(E) \,|\, \theta_1^\circ, \theta_2\} \equiv P\{\delta(E) \subset \theta_1^\circ \,|\, \theta_1^\circ, \theta_2\}$$

$$\equiv P\{E \in A(\theta_1^\circ) \,|\, \theta_1^\circ, \theta_2\} \equiv \alpha \tag{7}$$

whatever be θ_1° and whatever be θ_2.

Now what does this equation say? It says that if θ_1° be the true value of the parameter θ_1, then, whatever be the true value of θ_2, the probability of E falling within the region of acceptance $A(\theta_1^\circ)$ corresponding to θ_1° is equal to α.

What is the conclusion? If the functions $\underline{\theta}$ and $\overline{\theta}$ satisfy the condition (1) that we wanted, then the region of acceptance $A(\theta_1^\circ)$ corresponding to any possible value of θ_1 must have the property that the probability of the sample point E falling within this region, calculated under the assumption that θ_1° is the true value of θ_1, is independent of the value of θ_2 and equal to α. On every plane $\theta_1 = \text{const.}$ there will be a region of acceptance, and all such regions satisfy the condition stated.

This is one of the necessary conditions, say (i), which the regions of acceptance must satisfy if the functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ do possess the property stated by Eq.(1). There are a few others referring not to any particular region of acceptance, but the whole system of them.

(ii) Whatever be the sample point E, there must exist at least one value θ_1' of θ_1 such that $E \in A(\theta_1')$.

(iii) If $\theta_1' < \theta_1''$ and a sample point E falls within both $A(\theta_1')$ and $A(\theta_1'')$, then it must fall within any region of acceptance $A(\theta_1''')$ for which $\theta_1' < \theta_1''' < \theta_1''$.

(iv) If $\theta_1' < \theta_1''$, and if the sample point E falls within any one of the regions of acceptance $A(\theta_1)$ corresponding to $\theta_1' < \theta_1 < \theta_1''$, then it must fall within both $A(\theta_1')$ and $A(\theta_1'')$, and also within all other regions of acceptance $A(\theta_1)$ for which $\theta_1' < \theta_1 < \theta_1''$.

I have no time to give you the proofs,* but if the functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ possess the property described in Eq.(1), then the regions of acceptance $A(\theta_1)$ must possess the properties (i) - (iv). It is easy to see that this result reduces the construction of the upper and lower estimates of θ_1 to the problem of determining on each of the planes $\theta_1 = \theta_1'$ a region $A(\theta_1')$ such that each of them separately possesses the property (i) and their system has the properties (ii) - (iv). Again I must leave this without proof.

However, I will indicate how the functions $\underline{\theta}$ and $\overline{\theta}$ are defined, once the regions $A(\theta_1)$ satisfying (i) - (iv) have been determined. Take any sample point E. According to (ii) above there will be at least one value θ_1' of θ_1 such that $E \in A(\theta_1')$. Now take the maximum of the values of θ_1 for which $E \in A(\theta_1)$ and denote it by $\overline{\theta}(E)$. Likewise the minimum value of θ_1 for which $E \in A(\theta_1)$ will be denoted by $\underline{\theta}(E)$. It is not difficult to see that $\underline{\theta}(E)$ and $\overline{\theta}(E)$ thus defined do possess all the properties of the lower and upper estimates of θ_1.

---

* For details see my paper cited on page 28.

DR. DEMING: The system of those regions of acceptance will form some sort of tube.

DR. NEYMAN: That is right. This tube may be more or less complicated. Now that the problem of confidence intervals is reduced to that of the regions of acceptance, we may look into the question whether this is easily solved.

I have started with the assumption that the elementary probability law of the x_i depends on two parameters θ_1 and θ_2, of which we desire to estimate only the first. If there were only one unknown parameter θ_1, namely, the one to be estimated, then the only difference in our discussion would refer to condition (i), in formulating it we should drop the words "whatever be the value of θ_2." It will be seen that in such a situation the solution of the problem is extremely easy and that there are millions of different systems of regions of acceptance satisfying (i) - (iv), and therefore many different functions $\underline{\theta}(E) \leq \overline{\theta}(E)$ possessing the properties of the lower and upper estimates. All we want to do is to select in the sample space W, for any fixed value of θ_1, say θ_1', a region $A(\theta_1')$ satisfying the condition that the integral of $p(E|\theta_1')$ over this region is equal to α, and perhaps to shift these regions a little so as to satisfy the conditions (ii) - (iv).

If however the probability law of the x_i depends not only on θ_1 but also on some other parameter θ_2, (and perhaps even a third, or more) then the situation becomes more complicated, because of the difficulty to satisfy property (i) identically, whatever the value of θ_2 may be. Here we come to the necessity of considering regions, called "similar to the sample space with regard to the parameter θ_2." The problem of such regions have been discussed* ** and in certain cases we know the solution. In these cases, and they are sufficiently broad, we are in position to construct the confidence intervals for the parameter θ_1. Further progress depends on that in the problem of similar regions.

Let us now consider very briefly the question of the choice between all possible systems of confidence intervals corresponding to the same confidence coefficient α. This choice is somewhat analogous to the choice between several games of chance, in which the probabilities of winning are all the same, but the sums to be won different. In the study of confidence intervals, however, the choice is a little more difficult. We should naturally try to get confidence intervals as short as possible, and it may seem that the problem is that of finding the pair of functions $\underline{\theta}$ and $\overline{\theta}$ for which the difference

$$\overline{\theta} - \underline{\theta}$$

is a minimum. This problem, however, does not have any solution.

---

* J. Neyman and E. S. Pearson, footnote on page 75.

** J. Neyman, footnote on page 121.

$\underline{\theta}$ and $\overline{\theta}$ are functions of the sample point, and if we take into consideration one sample in particular, say E', then it is possible to find the system of the upper and lower estimate of θ_1 such that the difference between them,

$$\overline{\theta}(E') - \underline{\theta}(E')$$

at that particular point E' is a minimum. But then the difference of the same functions calculated from other sample points will be enormous. This is a difficulty, and all we can do is to try to have the difference $\overline{\theta}(E) - \underline{\theta}(E)$ a minimum "in general." This could be defined starting with the following considerations.

If θ_1° denotes the true value of θ_1, then, for any of the possible systems of confidence intervals we must have

$$P\{\delta(E) \ C \ \theta_1^\circ \,|\, \theta_1^\circ, \theta_2\} \equiv \alpha \quad \text{as written in Eq.(7)}$$

Suppose now that θ_1' is not the true one, i.e. $\theta_1' \neq \theta_1^\circ$. In this circumstance it is obviously useless for the confidence interval to cover θ_1'; on the contrary, we may consider it to be an advantage if the confidence interval fails to cover θ_1', provided of course it does cover θ_1°--the true value. If a confidence interval covering the true value of θ_1 also covers θ_1', which is not the true value, then we may say that this interval is "too broad."

Starting from this remark, we may define the "shortest" system of confidence intervals corresponding to the confidence coefficient α as the one by which any value of θ_1 that is not the true one is covered with the smallest possible frequency; that is,

$$P\{\delta(E) \ C \ \theta_1' \,|\, \theta_1^\circ, \theta_2\} = \text{a minimum if } \theta_1' \neq \theta_1^\circ \qquad (8)$$

Owing to the identity (6), this definition is immediately reduced to the following conditions concerning the regions of acceptance:

$$P\{E \in A(\theta_1) \,|\, \theta_1, \theta_2\} \equiv \alpha \qquad (9)$$

$$P\{E \in A(\theta_1) \,|\, \theta_1', \theta_2\} \leq P\{E \in A(\theta_1') \,|\, \theta_1', \theta_2\} \qquad (10)$$

for any $\theta_1 \neq \theta_1'$ and for any region $A(\theta_1)$ satisfying (9).

You probably recognize that in the theory of testing statistical hypotheses we have a similar though a little less complicated problem of the so-called uniformly most powerful tests.

Unfortunately, the "shortest" system of confidence intervals does not always exist, but this is a situation familiar in mathematics.

We all know that frequently rational numbers representing a square root of a given number do not exist. So it is with real solutions of a quadratic, etc. If a particular problem has no solution, we have to formulate some other problem that would satisfy the practical statistician. So, whenever the "shortest" system of confidence intervals does not exist, we may look for the so-called "short unbiased" system, which is defined as follows.

Take the probability

$$P\{\delta(E) \supset \theta_1' | \theta_1, \theta_2\} \qquad (11)$$

of the confidence interval $\delta(E)$ covering some particular value θ_1' of θ_1. This is, of course, a function of θ_1. If $\theta_1 = \theta_1'$ then, according to the general properties of the confidence intervals, particularly as expressed by Eq.(7), this probability will be equal to α; that is to say, the relative frequency of $\delta(E)$ covering θ_1', when θ_1' happens to be the true value, will be α. Now if the true value of θ_1 is shifted either way from θ_1', we shall require that the probability (11) should fall off as quickly as possible. Altogether, then, we have these requirements:

$$\left.\begin{aligned} P\{\delta(E) \supset \theta_1' | \theta_1, \theta_2\} &= \alpha \\ \frac{d}{d\theta_1} P\{\delta(E) \supset \theta_1' | \theta_1, \theta_2\} &= 0 \\ \frac{d^2}{d\theta_1^2} P\{\delta(E) \supset \theta_1' | \theta_1, \theta_2\} &= \text{minimum} \end{aligned}\right\} \text{if } \theta_1 = \theta_1' \qquad (12)$$

As a matter of fact, the minimum value of the second derivative in (12) is always negative. Those conditions also are readily reduced to similar ones referring to the regions of acceptance.

If we manage to obtain a system of confidence intervals satisfying the conditions (12), for any possible value θ_1' of θ_1, then we shall call it short unbiased. The justification of the term unbiased is that if the conditions (12) are satisfied, and the second derivative is negative, then the true value of θ_1 will always be covered by the confidence interval more frequently than any other value.

All those points are not very familiar and I shall probably have to refer to my publication quoted (footnote page 28) where you will find many details and illustrations.

QUESTION: Did you say that the 2d derivative in (12) is always negative, whether you choose the minimum value or not, whether you get a minimum or not?

DR. NEYMAN: This minimum 2d derivative is always negative.

QUESTION: But did you say all values of the 2d derivative will be negative?

DR. NEYMAN: Whatever problem we consider, and whatever be the fixed value θ', we may find many systems of confidence intervals for which the first two conditions of (12) are satisfied, namely,

$$P\{\delta(E) \text{ C } \theta_1' | \theta_1', \theta_2\} = \alpha \quad \text{[as in the identity (9)]}$$

and

$$\frac{d}{d\theta_1} P\{\delta(E) \text{ C } \theta_1' | \theta_1, \theta_2\} = 0 \quad \text{for } \theta_1 = \theta' \quad \text{[the second condition in (12)]}$$

But for those systems, the second derivative

$$\frac{d^2}{d\theta_1^2} P\{\delta(E) \text{ C } \theta_1' | \theta_1, \theta_2\}$$

may have various values, and some of them will be positive; but my point is that if we choose the system for which this 2d derivative is smaller than for any other system, then the minimum value of the 2d derivative will be negative.

DR. DEMING: The question, if I might restate it, is why aren't you satisfied merely to have the 2d derivative negative? Why do you require it to be as small as possible?

DR. NEYMAN: This has something to do with the other part of the term used to describe the system satisfying the conditions in (12), namely, "short."

Let us look at what could be the graphs of

$$P\{\delta(E) \text{ C } \theta_1' | \theta_1, \theta_2\} \qquad (11)$$

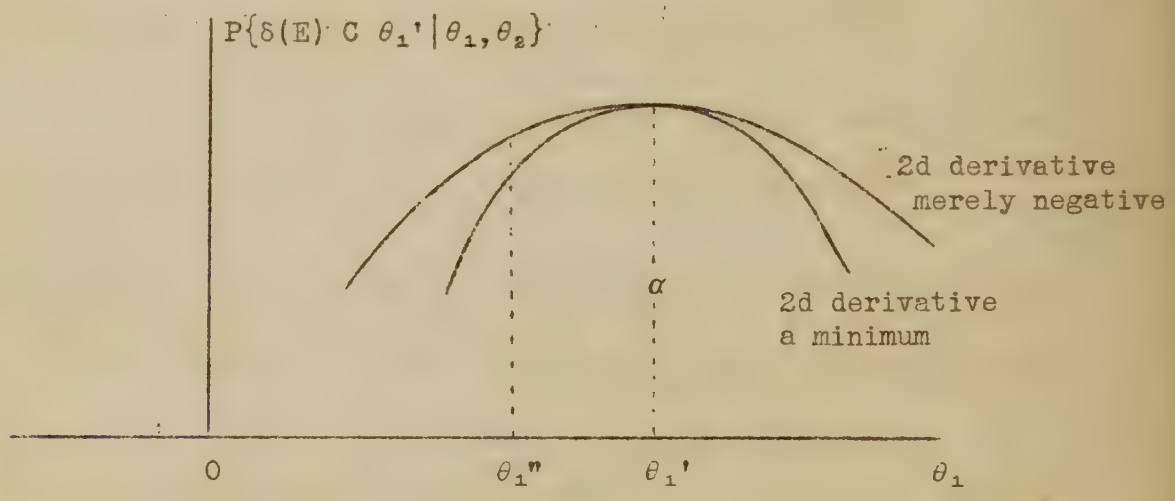
considered as functions of θ_1, corresponding to the two cases, when the second derivative in (12) is merely negative, and when it is a minimum. We shall get a picture like the one on the diagram on the next page. The two curves have the same ordinate α for $\theta_1 = \theta_1'$, and both have a maximum at that point, but the curve corresponding to the minimum of the second derivative will fall off quicker than the other. The smaller the 2d derivative, the steeper the maximum. In consequence, if the true value of θ_1 happens to be $\theta_1'' \neq \theta_1'$ then θ_1' will be less frequently covered by the confidence intervals if the condition of the minimum 2d derivative is satisfied. Of course, strictly speaking, this refers to values θ_1'' in the vicinity of θ_1'.

It may be useful to conclude this exposition by quoting a

practical example. Consider the case where it is known that all the observations $x_1, x_2, \ldots, x_n$ follow the same normal law and are mutually independent. Then

$$p(E|\mu,\sigma) = [\sigma\sqrt{(2\pi)}]^{-n} e^{-\Sigma (x_i - \mu)/2\sigma^2} \qquad (13)$$

where μ and σ are unknown. Suppose it is desired to estimate μ. This is just the case where the probability law depends on two unknown parameters and we have to deal with regions similar to the sample space with respect to σ.



The short unbiased system of confidence intervals is provided by

$$\begin{aligned} \overline{\mu}(E) &= \overline{x} + t_\alpha \, s/\sqrt{(n-1)} = \overline{x} + t_\alpha \, s'/\sqrt{n} \\ \underline{\mu}(E) &= \overline{x} - t_\alpha \, s/\sqrt{(n-1)} = \overline{x} - t_\alpha \, s'/\sqrt{n} \end{aligned} \qquad (14)$$

where $\overline{x}$ is the sample mean, s the S.D. of the sample defined by

$$s^2 = \frac{1}{n} \Sigma (x_i - \overline{x})^2 \qquad (15)$$

as was used on page 134. s' is the estimate of σ written on page 135; and t_α may be taken directly from Fisher's tables, according to the number of degrees of freedom n-1 and corresponding to his $P = 1-\alpha$.

You see that in this particular example the theory of estimation is not dogmatic, and that it brings us to a solution of the problem equivalent to what is familiar. This is an illustration of my statement at the end of the preceding conference to the effect that the rôle of the theory is frequently very modest compared to the pioneer achievements of the practical man.

However, it is worth noting that the traditional procedure of estimating σ in the same case (13) is less successful: the corresponding confidence intervals are biased and the frequency of their covering some of the wrong values of σ is actually greater than that of covering the true one.* But even here the advantage of the unbiased system of confidence intervals is only a very slight one: what the uncontrollable instinct of the practical man has overlooked is relatively unimportant.

MRS. KANTOR: Is there anything in English literature on confidence intervals?

DR. NEYMAN: It is a relatively new subject. Apart from my paper in the Phil. Trans. already cited on page 28, where I give the theory as discussed here, the references are as follows.**

1. J. Neyman: On the two different aspects of the representative method. J. Roy. Stat. Soc. 97, 558-625, 1934. See particularly pp. 589-593. You will find here the description of the general idea in the simplest case. Formula (5) on p.565 (without proof) is an analogue of my formulas (14) here. Cited on page 90.

2. C. J. Clopper and E. S. Pearson: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26 404-413, 1934.

3. T. Matuszewski, J. Neyman, and J. Supinska: Statistical studies in questions of bacteriology. Supplement J. Roy. Stat. Soc. 2, 63-82, 1935. Here are given tables of the confidence intervals for the concentration of living bacteria in a suspension.

4. J. Neyman: On the problem of confidence intervals. Am. Math. Statistics, 6, 111-116, 1935. Here it is shown that in the case when the x_i are discontinuous, it may be impossible to satisfy the condition (i) exactly, but only in the form of an inequality $P\{\delta(E) \subset \theta_1^0 | \theta_1^0\} \geq \alpha$.

5. J. Przyborowski and H. Wilenski: Statistical principles of routine work in testing clover seed for dodder. Biometrika, 27, 273-292, 1935. Here the authors give the confidence intervals for the unique constant on which the Poisson distribution depends. Cited on page 27.

---

* A similar circumstance exists in the problem of testing hypotheses when standard deviations are involved, as was brought out by E. S. Pearson in the discussion of a paper presented by R. A. Fisher at a meeting of the Royal Statistical Society in December 1934. See the J. Roy. Stat. Soc. 98, 39-82, 1935. pages 66-67 in particular. Editor.

** The notion underlying confidence intervals seems to have been introduced by E. B. Wilson, J. Amer. Stat. Assoc. 22, 209-212, 1927. Editor.

6. Robert W. B. Jackson: Some problems of testing statistical hypotheses and estimation relating to the question of the relative accuracy of measurement. (To be published in the Statistical Research Memoirs, 2, 1937). The author applies the general theory as sketched in these conferences to derive the confidence intervals relating to various psychological problems.

With respect to all the previous publications concerning confidence intervals for which I am either totally or partly responsible, I have to say that they contain a certain artificiality which is now removed and of which the last publication in the Philosophical Transactions 1937 is free (footnote page 28). This artificiality consisted in assuming that the unknown parameter to be estimated is a random variable itself following an arbitrary probability law $p(\theta)$. The arbitrariness of $p(\theta)$ extended to the situation where it could reduce to unity just for one particular value θ° of θ, being zero elsewhere, in which case θ would be a constant. This circumstance served as an excuse, but the mere assumption of θ being a random variable does seem to be artificial.

I should perhaps make a second remark concerning the connection between confidence intervals and the fiducial probability theory of R. A. Fisher. In my paper of 1934 I stated, as I thought then, that the two theories are essentially the same. However I made some protest against the terms "fiducial probability" and "fiducial distribution" of the parameter θ, for which Fisher* adopted the notation

$$y\, d\theta = f(T|\theta)\, d\theta \qquad (16)$$

T being some function of the x_i. Later, however, when discussing my paper, Fisher insisted that "fiducial distribution" is a conception essential in his theory. Similar statements regarding fiducial distributions of parameters you will find in his recent article in the Annals of Eugenics, together with some criticism of the theory of confidence intervals.**

DR. DEMING: There was also a publication by T. E. Sterne in the Proceedings of the National Academy, very much of the same thing.

DR. NEYMAN: These statements of Fisher forced me to alter my opinion, and now I think that the theory of confidence intervals and of

---

* R. A. Fisher, "Inverse probability," Proc. Cambridge Phil. Soc. 26, 528-535, 1930; page 534 in particular.

** R. A. Fisher, "The fiducial argument in statistical inference," Annals of Eugenics, 6, 391-398, 1936.

T. E. Sterne, Proc. Nat'l Academy Sci. 20, 601-603, 1934.

fiducial probability are two different things.

DR. DEMING: Fisher uses his "fiducial distributions" to calculate fiducial limits, in other words, confidence limits. His "fiducial distribution" is not used in differential form, but in integrations, as I understand it.

DR. NEYMAN: That is right, and the numerical results are the same. It is possible that they will be the same always. But the theories seem to differ. In the theory of confidence intervals there is no room for anything like the fiducial distribution of θ in the form (16). My impression is that the difference between the two theories is recognized also by Fisher: in his last publication in the Annals of Eugenics (footnote page 158) he warns the reader of some contradictions allegedly inherent in the theory of confidence intervals.

DR. DEMING: Could not one very easily dispose of the notion of a "fiducial distribution of θ" by simply pointing out that if in repeated sampling you were to make up a frequency table of the various values of θ that occur in samples, all selected for a particular value or differential range of T, it would be found that the distribution of θ is what is commonly called the prior distribution of θ, and this may be and usually would be entirely unrelated to the fiducial distribution (16)?

DR. NEYMAN: I must admit that I am not able to follow Fisher's theory and therefore I cannot criticize. I know only that in the theory of confidence intervals there is no room for anything like $d\theta$.

MR. WILLIAM C. SHELTON: In solving algebraic equations, for example, you may say you have an approximation; then you say you will improve upon this approximation. One way of doing so is to use Newton's method of approximation. In that case we do differentiate with respect to a constant. We do it simply because we realize--we know it has been proved--that in order to find the value of the unknown, we may consider it as a variable and locate the root of the equation. We are not assuming that the unknown that we are trying to find actually varies.

DR. NEYMAN: My impression is that the situation is somewhat different. Formerly I had thought something similar to what you say, and I said (in 1934) that "fiducial probability" and "fiducial distribution" are faulty terms, sort of suggesting conceptions that were not in the mind of their author. I find however that this is not so. For look what R. A. Fisher said himself when discussing my paper of 1934 (p.618 loc. cit.) "Here, again, there might be serious difficulties in respect to the mutual consistency of the different inferences to be drawn; for, with a single parameter, it could be shown that all the inferences might be summarized in a single probability distribution for

that parameter, and that, for this reason, all were mutually consistent;..." I have underlined the passage that seems to be relevant, and you will notice that even the qualifying "fiducial" has been left out. But I will repeat again that I do not understand the fiducial theory and, in particular, I am not quite clear what R. A. Fisher had in mind when writing the above statements, or what are the possible inconsistencies he mentioned.

DR. DEMING: This second derivative being necessarily not only negative but as small a negative number as possible, is this not tied up directly with your idea of trying to avoid "errors of the second kind?"

DR. NEYMAN: Both in the theory of estimation and in that of testing hypotheses we have many similar formulas and, in fact, many similar ideas. However, the conception of errors of the first and second kinds (p.45) is specific to the theory of testing hypotheses and does not enter into that of estimation. The reason is that as a result of testing a statistical hypothesis, two kinds of action are possible: (1) reject the hypothesis, and (2) do not reject the hypothesis. In consequence there are two different ways in which we may be wrong, i.e. there are two kinds of errors to be cautious of. On the other hand, in the theory of estimation the result is always of the same form: we say θ is this or that, satisfies this inequality or some other. This statement may be correct or wrong, but there is only one way of its being wrong: θ is not what it is stated to be.

INDEX

INDEX

INDEX