Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability
Author(s): J. Neyman
Source: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, Vol. 236, No. 767 (Aug. 30, 1937), pp. 333-380
Published by: The Royal Society
Stable URL: http://www.jstor.org/stable/91337
Accessed: 09/12/2010 08:43

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. NEYMAN

Reader in Statistics, University College, London

(Communicated by H. JEFFREYS, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

CONTENTS

I—INTRODUCTORY

(*a*) *General Remarks, Notation, and Definitions*

We shall distinguish two aspects of the problems of estimation : (i) the practical and (ii) the theoretical. The practical aspect may be described as follows :

(i*a*) The statistician is concerned with a population, π, which for some reason or other cannot be studied exhaustively. It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population π. For example, it may be desired to calculate approximately the mean of a certain character possessed by the individuals forming the population π, etc.

(i*b*) Alternatively, the statistician may be concerned with certain experiments which, if repeated under apparently identical conditions, yield varying results. Such experiments are called random experiments, (*see* p. 338). To explain or describe

the machinery of the varying results of random experiments certain mathematical schemes are drawn up involving one or more parameters, the values of which are not fixed. The statistician is then asked to provide numerical values of these parameters, to be calculated from experimental data and upon the assumption that the mathematical model of the experiments is correct.

The situation may be exemplified by the counts of α-particles ejected by some radioactive matter. The physicists have here elaborated a mathematical model of the phenomenon involving only one numerical parameter, namely, the average duration of life of an atom, and the statistician is asked to use the results of the available observations to deduce the numerical value of this parameter.

In both cases described, the problem with which the statistician is faced is the problem of estimation. This problem consists in determining what arithmetical operations should be performed on the observational data in order to obtain a result, to be called an estimate, which presumably does not differ very much from the true value of the numerical character, either of the population π, as in (i*a*), or of the random experiments, as in (i*b*).

(ii) The theoretical aspect of the problem of statistical estimation consists primarily in putting in a precise form certain vague notions mentioned in (i). It will be noticed that the problem in its practical aspect is not a mathematical problem, and before attempting any mathematical solution we must substitute for (i) another problem, (ii), having a mathematical sense and such that, for practical purposes, it may be considered as equivalent to (i).

The vague non-mathematical elements in (i) are connected with the sentence describing the meaning of the word estimate. What exactly is meant by the statement that the value of the estimate " presumably " should not differ very much from the estimated number ? The only established branch of mathematics dealing with conceptions bearing on the word " presumably " is the calculus of probability. It therefore seems natural to base the precise definition of an estimate on conceptions of probability. It is easy to see that the connexion of the problem considered with the theory of probability does not stop here and that the conditions of the problem themselves are, mathematically, clear only if they are expressed in the same terms of probability.

In (i*a*) we speak of a statistician drawing a sample from the population studied. It is known that if the sample is systematically selected and not drawn " at random " the conclusions concerning the population π formed on its basis are, as a rule, false and at the present state of our knowledge impossible to justify. On the other hand, we know that justifiable and frequently correct conclusions are possible only when the process of drawing the sample is " random ", though the randomness may be at times more or less restricted. I have put the word " random " in inverted commas because it is very difficult to define what is meant by it in practice.* We try to achieve randomness by more or less complicated devices, using roulette,

* This point requires a longer discussion, which I hope to be able to publish in a separate paper.

dice, etc. Theoretically, however, the situation is clear : when we speak of a random sample we mean that it is drawn so that (1) the probability of each individual of the population being included in the sample is the same, and (2) separate drawings are mutually independent, except in the case of dependence resulting from the population being finite, when the individual drawn is not returned to the population before the next drawing.

Leaving apart on one side the practical difficulty of achieving randomness and the meaning of this word when applied to actual experiments, I want to call attention to the fact that the conditions of the problem in (i*a*) may be mathematically described as follows.

Denote X, Y, . . ., Z, the characters of the individuals of the population π, in which we are interested and by $x, y, \ldots . z$, respectively the values of these characters corresponding to some particular individual. For example, if the population π consists of certain plants, X may mean the weight of the roots, Y the colour of the flowers, Z the weight of the seeds, etc. The method of random sampling adopted, together with the properties of the population π, some of which may be known and others doubtful, determine the probability,* say P {E}, of the occurrence of any possible system, E, of values of X, Y, ..., Z in the individuals which may be drawn to form the sample. Denote by θ_1 the numerical character of the population π which it is desired to estimate : this, for example, may be the mean value of X, the regression coefficient of Z on X, the mean square contingency of Z and Y, etc. The probability P {E} will depend on the value of θ_1 and in most cases on the values of certain other parameters, say, θ_2, θ_3, . . ., etc.

We see, therefore, that the problem with which the theoretical statistician is faced is as follows :

Sampling randomly from the population π, it is possible to obtain samples, say

$$E_1, E_2, \ldots, E_N, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where each sample is described by means of values of the characters X, Y, . . ., Z, corresponding to each of the individuals forming the sample. The probability of any sample E_i, say $P\{E_i | \theta_1, \theta_2, \ldots \theta_l\}$, depends on a certain number, l, of parameters θ_i, the values of which are unknown, describing the properties of the population π. The problem consists in determining how to use the sample which may be actually obtained in order to estimate θ_1.

We see that the conditions of the problem in (i*a*) are expressed in terms of probability. The same holds good with regard to the problem in (i*b*), which shows that the distinction between (i*a*) and (i*b*) is only superficial. In fact, random experiments differ from those which are not considered as random only by the circumstance that the mathematical model devised for their description involves

* If the population π is finite. Otherwise the method of sampling and the properties of the population will determine the elementary probability law of X, Y, ..., Z considered as random variables. For the definitions of random variables and their probability laws, *see* p. 340 below.

probabilities. Each model of this kind determines the range of the possible results of random experiments and also the probability of each such result, depending upon one parameter or more, the numerical value of which is unknown.

We come to the conclusion that both the conditions of the problem of estimation and the satisfactory solution sought, if expressed accurately, are expressed in terms of probability. Before we proceed to the final formulation of the problem, it will be useful to give a short review of the forms of some solutions which have been advanced in the past. For this we shall need to define the terms probability, random variable, and probability law. These definitions are needed not because I introduce some new conceptions to be described by the above terms, but because the theory which is developed below refers only to some particular systems of the theory of probability which at the present time exist,* and it is essential to avoid misunderstandings.

I find it convenient to use the word probability in the following connexion : " the probability of an object, A, having a property B ". This may include as particular cases : " probability of a result, A, of a certain experiment having the property B of actually occurring " (= probability of the result A — for short) and " the probability of a proposition, A, of having the property, B, of being true ". All these ways of speaking could be shortened in obvious ways.

I want to emphasize at the outset that the definition of probability as given below is applicable only to certain objects A and to certain of their properties B—not to all possible. In order to specify the conditions of the applicability of the definition of the probability, denote by (A) the set of all objects which we agree to denote by A. (A) will be called the fundamental probability set. Further, let (B) denote the set of these objects A which possess some distinctive property B and finally, ((B)), a certain class of subsets (B') , (B''), . . ., corresponding to some class of properties B', B'', etc.

It will be assumed†

(1) that the class ((B)) includes (A), so that (A) ε ((B)) and

* It may be useful to point out that although we are frequently witnessing controversies in which authors try to defend one or another system of the theory of probability as the only legitimate, I am of the opinion that several such theories may be and actually are legitimate, in spite of their occasionally contradicting one another. Each of these theories is based on some system of postulates, and so long as the postulates forming one particular system do not contradict each other and are sufficient to construct a theory, this is as legitimate as any other. In this, of course, the theories of probability are not in any sort exceptional.

Both Euclidean and non-Euclidean geometries are equally legitimate, but, *e.g.*, the statement " the sum of angles in a linear triangle is always equal to π " is correct only in the former. In theoretical work the choice between several equally legitimate theories is a matter of personal taste only. In problems of application the personal taste is again the decisive moment, but it is certainly influenced by considerations of the relative convenience and the empirical facts.

† The problem of the definition of measure in relation to the theory of probability has been recently discussed by Łomnicki and Ulam (1934), who quote an extensive literature. A systematic outline of the theory of probability based on that of measure is given by Kolmogoroff (1933). *See also* Borel (1925–26) ; Lévy (1925) ; Fréchet (1937).

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

(*a*) Whatever (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 neither pair (B_i) and (B_j) (where $i \neq j$) have common elements, then m (ΣB_i) $= \sum_{i=1}^{\infty} m$ (B_i).

(*e*) If (B) is measurable, then the set ($\overline{B}$) of objects A non-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 $P\{B|A\} = \frac{m\ (B)}{m\ (A)}$. The probability P{$B_1$|A}, or P{$B_1$} for short, may be called the absolute probability of the property B_1. Denote by B_1 B_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_1 B_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) $= 0$ and we are able to define some other measure, say m', applicable to (B_2) and to a class of its subsets including (B_1 B_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 $P\{B_1B_2\} = P\{B_1\}P\{B_2|B_1\} = P\{B_2\}P\{B_1|B_2\}$.

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-dimensioned space, then the measure of Lebesgue will satisfy the definition used here. These two definitions will be used wherever applicable.

If (A) is infinite but the objects A are not actually points (*e.g.*, if they are certain lines, etc.), the above definition 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 historical example considered by BERTRAND, POINCARÉ, and BOREL when the objects A are the chords in a circle C of radius r and the property B consists of their length, l, exceeding some specified value, B. It may be useful to consider two of the possible ways of treating this problem.

1. Denote by x the angle between the radius perpendicular to any given chord A and any fixed direction. Further, let y be the distance of the chord A from the centre of the circle C. If A′ denotes a point on the plane with coordinates x and y, then there will be a one to one correspondence between the chords A of length $0 \leqq l < 2r$ and the points of a rectangle, say (A′), defined by the inequalities $0 < x \leqq 2\pi$ and $0 < y \leqq r$. The measure of the set of chords A with lengths exceeding B could be defined as being equal to the area of that part of (A′) where $0 < y \leqq \sqrt{r^2 - (\frac{1}{2}\mathrm{B})^2}$. It follows that the probability in which we are interested is $\mathrm{P}\{l > \mathrm{B}\} = (r^2 - (\frac{1}{2}\mathrm{B})^2)^{\frac{1}{2}} r^{-1}$.

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 A″ within the parallelogram (A″) determined by the inequalities $0 < x \leqq 2\pi$, $x \leqq y \leqq x + \pi$. The measure of the set of chords A with their lengths exceeding B may be defined as being equal to the area of that part of (A″) where $2r \sin \frac{1}{2} y > \mathrm{B}$.

Starting with this definition $\mathrm{P}\{l > \mathrm{B}\} = 1 - 2 \arcsin (\mathrm{B}/2r)\, \pi^{-1}$.

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 they correspond to different conditions of the problem. In fact, the question " what is the probability of a chord having its length larger than B " 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 solution (1), we could say that when selecting at random a chord A, we first pick up at random the direction of a radius, all of them 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 solution (2). We shall use sometimes 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 this paper the sentence like the one taken in inverted commas 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 here, and it is a pure convention that the statement

" In picking up at random a chord, we first select a direction of radius, all of them 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_1) of chords is defined as that of the set (A'_1) of points with coordinates x and y such that for any chord A_1 in (A_1), x is the direction of the radius perpendicular to A_1 and y the distance of A_1 from the centre of the circle. (A_1) is measurable only if (A'_1) is so."

However free we are in mathematical work in using wordings we find convenient, as 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 on calling the law of big numbers. This is 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 (BORTKIEWICZ, 1917), 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 Tables of random numbers (1927).

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, say E, 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 (A) adequately represents the method of carrying out the experiment E. The theory developed below is entirely independent of whether the law of big numbers holds

* This is mentioned by BOREL (1910). I could not find the name of the performer of the experiment.

good or not. But the applications of the theory do depend on the assumption that it is valid. The questions dealt with in the present section are of fundamental importance. However, they do not constitute the main part of the paper and therefore are necessarily treated very briefly. The readers who may find the present exposition not sufficiently clear may be referred for further details to the work of KOLMOGOROFF (1933, *see* particularly p. 3 *et seq.*). I should state also that an excellent theoretical explanation of the experimental phenomena mentioned, connected with the previous work of POINCARÉ and SMOLUCHOWSKI, has been recently advanced by HOPF (1934).

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 and also its relative probability given any other property B_1, must be equal to unity, so that $P\{B_0\} = P\{B_0|B_1\} = 1$. On the other hand, if it is known only that $P\{B_0\} = 1$, then it does not necessarily follow that $P\{B_0|B_1\}$ must be equal to unity.

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

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

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

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

A random variable is here contrasted with a constant, say θ, which will be defined as a magnitude, 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 shall often use the way of speaking 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 $\pi = 3{\cdot}14159\ldots$ 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 that X_{1000} may have does not depend upon the result of any random experiment.

Similarly, if we consider a specified population, say the population π_{1935} of persons residing permanently in London during the year 1935, any character of this population will be a constant. In the sense of the terms used here, there will be no practical meaning in a question concerning the probability that the average income, say I_{1935}, of the individuals of this population is, say, between £100 and £300. As the fundamental probability set consists of only one element, namely I_{1935}, the value of this probability is zero or unity, and to ascertain it we must discover for certain whether $£100 \leq I_{1935} < £300$ or not. This is, of course, possible, though it might involve great practical difficulty, just as it is possible to find the actual value of X_{1000}, the 1000th figure in the expansion of π. Any calculations showing that $P\{100 \leq I_{1935} < 300\}$ has a greater value than zero and smaller than unity must be either wrong or based on some theory of probability other than the one considered here.

This is the point where the difference between the theory of probability adopted here and that developed by JEFFREYS (1931) comes to the front. According to the latter, previous economic knowledge may be used to calculate the probability $P\{a \leq I_{1935} < b|B\}$ where $a < b$ are any numbers and the result of the calculations may be represented by any fraction, not necessarily by zero or unity.

The above examples must be contrasted with the following ones. We may consider the probability of a figure X, in the expansion of π falling between any specified limits $a < b$ and find it to be equal, *e.g.*, to $\frac{1}{2}$. This is possible when we first define a random method of drawing a figure out of those which serve to represent the expansion of π. If this is done, then X is a random variable and the X_{1000} previously defined will be one of its particular values.

Similarly, it is probably not impossible to construct a more or less adequate mathematical model of fluctuations in the size of income, in which the yearly average income, I, of the permanent population of London will be a random variable. The I_{1935} previously defined will be a particular value of I, observed at the end of the year 1935.

It is true that any constant, ξ, might be formally considered as a random variable

with the integral probability law $P\{a \leq \xi < b\}$ having only values unity or zero according to whether ξ falls between a and b or not. If we pass from letters to figures this will lead to formulae like $P\{1 \leq 2 < 3\} = 1$, or $P\{3 \leq 2 < 4\} = 0$.

Of course, in practice we shall have generally some unknown number ξ instead of 2 in the above formulae and accordingly we shall not know what are the actual values of the probabilities. In order to find these values, it would be necessary to obtain some precise information as to the value of ξ. It follows that the consideration of such probabilities is entirely useless, since whatever we are able to express in using them, we can say more simply by means of equations or inequalities.

For this reason, when defining a random variable, we require its probability law to be able to have values other than zero and unity. The other case may be set aside as trivial.

In the following development we shall have to consider at once several random variables

$$X_1, X_2, \ldots X_n. \quad \ldots\ldots\ldots\ldots \quad (2)$$

It will be convenient to denote by E any combination of their particular values and to interpret each such combination E as a point (the sample point) in an n-dimensional space W (the sample space), having its coordinates equal to the particular values of the variables (2). If w denotes any region in W, then the probability, say $P\{E \varepsilon w\}$, of the sample point falling within w considered as a function of w will be described as the integral probability law of the variables (2).

We shall consider only cases where there exists a non-negative function $p(E) \equiv p(x_1, \ldots, x_n)$ determined and integrable in the whole sample space W, such that for any region w

$$P\{E \varepsilon w\} = \int \ldots \int_w p(E)\, dx_1 \ldots dx_n. \quad \ldots\ldots\ldots \quad (3)$$

The function $p(E)$ will be called the elementary probability law of the X's in (2). It is easy to show that when $p(x_1, \ldots, x_{n-1}, x_n)$ is known, then $p(x_1, \ldots, x_{n-1})$ may be calculated by integrating $p(x_1, \ldots x_n)$ with regard to x_n from $-\infty$ to $+\infty$.

When dealing with several probability laws calculated in relation to probability sets depending on some variables, say $y_1 \ldots y_m$, in order to avoid misunderstandings, we shall use the notation $p(x_1 \ldots x_n | y_1 \ldots y_m)$ or $p(E | y_1 \ldots y_m)$. If $p(x_1, \ldots x_k, x_{k+1}, \ldots x_n)$ is the probability law of $x_1, x_2 \ldots x_k, x_{k+1}, \ldots x_n$ and if for a given system of the x's, $p(x_{k+1}, \ldots x_n) > 0$ then, for that system, the relative probability law of $x_1, x_2, \ldots x_k$ given $x_{k+1}, \ldots x_n$, denoted by $p(x_1, \ldots x_k | x_{k+1}, \ldots x_n)$, will be defined by the relation $p(x_1, x_2, \ldots x_k, \ldots x_n) = p(x_{k+1}, \ldots x_n)\, p(x_1, \ldots x_k | x_{k+1}, \ldots x_n)$.

With the above definitions and notation we may now formulate the problem of estimation as follows :

Let

$$X_1, X_2, \ldots X_n \quad \ldots\ldots\ldots\ldots \quad (4)$$

be a system of n random variables, the particular values of which may be given by observation. The elementary probability law of these variables

$$p(x_1 \ldots x_n|\theta_1, \theta_2, \ldots \theta_l) \quad \ldots\ldots\ldots\ldots \quad (5)$$

depends in a known manner upon l parameters $\theta_1 \ldots \theta_l$, the values of which are not known. It is required to estimate one (or more) of these parameters, using the observed values of the variables (4), say

$$x'_1, x'_2, \ldots x'_n \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

(b) Review of the Solutions of the Problem of Estimation Advanced Hereto*

The first attempt to solve the problem of estimation is connected with the theorem of Bayes and is applicable when the parameters $\theta_1, \theta_2, \ldots \theta_l$ in (5) are themselves random variables. The theorem of Bayes leads to the formula

$$p(\theta_1, \theta_2, \ldots \theta_l|x'_1, x'_2, \ldots x'_n) = \frac{p(\theta_1, \theta_2, \ldots \theta_l)\, p(x'_1, x'_2, \ldots x'_n|\theta_1 \ldots \theta_l)}{\int \ldots \int p(\theta_1, \theta_2, \ldots \theta_l)\, p(x'_1, x'_2, \ldots x'_n|\theta_1, \ldots \theta_l)\, d\theta_1 \ldots d\theta_l}, \quad (7).$$

representing the probability law of $\theta_1, \theta_2, \ldots \theta_l$, calculated under the assumption that the observations have provided the values (6) of the variables (4). Here $p(\theta_1, \ldots \theta_l)$ denotes the probability law of the θ's, called *a priori*, and the integral in the denominator extends over all systems of values of the θ's. The function $p(\theta_1, \theta_2, \ldots \theta_l|x'_1, x'_2 \ldots x'_n)$ is called the *a posteriori* probability law of θ's. In cases where the *a priori* probability law $p(\theta_1, \theta_2, \ldots \theta_l)$ is known, the formula (7) permits the calculation of the most probable values of any of the θ's and also of the probability that θ_i, say, will fall in any given interval, say, $a \leq \theta_i < b$. The most probable value of θ_i, say $\overset{v}{\theta}_i$, may be considered as the estimate of θ_i and then the probability, say

$$P\{\overset{v}{\theta}_i - \Delta < \theta_i < \overset{v}{\theta}_i + \Delta|E'\}, \quad \ldots\ldots\ldots \quad (8)$$

will describe the accuracy of the estimate $\overset{v}{\theta}_i$, where Δ is any fixed positive number and E′ denotes the set (6) of observations.

It is known that, as far as we work with the conception of probability as adopted in this paper, the above theoretically perfect solution may be applied in practice only in quite exceptional cases, and this for two reasons :

(*a*) It is only very rarely that the parameters $\theta_1, \theta_2, \ldots \theta_l$ are random variables. They are generally unknown constants and therefore their probability law *a priori* has no meaning.

* This review is not in any sense complete. Its purpose is to exemplify the attempts to solve the problem of estimation.

(*b*) Even if the parameters to be estimated, $\theta_1, \theta_2, \ldots \theta_l$, could be considered as random variables, the elementary probability law *a priori*, $p(\theta_1, \theta_2, \ldots \theta_l)$, is usually unknown, and hence the formula (7) cannot be used because of the lack of the necessary data.

When these difficulties were noticed, attempts were made to avoid them by introducing some new principle lying essentially outside the domain of the objective theory of probability.

The first of the principles advanced involved the assumption that when we have no information as to the values of the θ's, it is admissible to substitute in formula (7) some function of the θ's selected on intuitive grounds, *e.g.*,

$$p(\theta_1, \theta_2, \ldots \theta_l) = \text{const.} \quad \ldots\ldots\ldots\ldots \quad (9)$$

and use the result, say

$$p_1(\theta_1, \ldots \theta_l | \mathrm{E}') = \frac{p(x'_1, x'_2, \ldots x'_n | \theta_1, \ldots \theta_l)}{\int \ldots \int p(x'_1, x'_2 \ldots x'_n | \theta_1, \ldots \theta_l)\, d\theta_1 \ldots d\theta_l}, \quad \ldots\ldots \quad (10)$$

as if this were the *a posteriori* probability law of the θ's.

This procedure is perfectly justifiable on the ground of certain theories of probability, *e.g.*, as developed by HAROLD JEFFREYS, but it is not justifiable on the ground of the theory of probability adopted in this paper. In fact, the function $p_1(\theta_1 \ldots \theta_l | \mathrm{E}')$ as defined by (10) will not generally have the property serving as a definition of the elementary probability law of the θ's. Its integral over any region w in the space of the θ's will not be necessarily equal to the ratio of the measures of two sets of elements belonging to the fundamental probability set, which we call the probability. Consequently, if the experiment leading to the set of values of the x's is repeated many times and if we select such experiments (many of them) in which the observed values were the same, $x'_1, x'_2 \ldots x'_n$, the assumed validity of the law of big numbers (in the sense of BORTKIEWICZ) will not guarantee that the frequency of cases where the true value of θ_i falls within $\overset{v}{\theta}_i - \Delta < \theta_i < \overset{v}{\theta}_i + \Delta$ will approach the value of (8), if this is calculated from (10). Moreover, if the θ's are constant, this frequency will be permanently zero or unity, thus essentially differing from (8).

The next principle I shall mention is that advocating the use of the so-called unbiassed estimates and leading to the method of least squares. Partly following MARKOFF (1923), I shall formulate it as follows :

In order to estimate a parameter θ_i involved in the probability law (5), we should use an unbiassed estimate or, preferably, the best unbiassed estimate.

A function, F_i, of the variables (4) is called an unbiassed estimate of θ_i if its mathematical expectation is identically equal to θ_i, whatever the actual values of $\theta_1, \theta_2, \ldots \theta_l$. Thus,

$$\mathscr{E}(\mathrm{F}_i) \equiv \theta_i. \quad \ldots\ldots\ldots\ldots \quad (11)$$

An unbiassed estimate F_i is called the best if its variance, say

$$V_{F_i} = \mathscr{E}(F_i - \theta_i)^2, \quad \ldots\ldots\ldots\ldots \quad (12)$$

does not exceed that of any other unbiassed estimate of θ_i.

It is known that MARKOFF provided a remarkable theorem leading, in certain cases, to the calculation of the best of the unbiassed estimates which are linear functions of the variables (4). The advantage of the unbiassed estimates and the justification of their use lies in the fact that in cases frequently met the probability of their differing very much from the estimated parameters is small.

The other principle, which is to a certain extent in rivalry with that of the unbiassed estimate, is the principle of maximum likelihood. This consists in considering $L = \text{const.} \times p(x'_1, x'_2 \ldots x'_n | \theta_1 \ldots \theta_l)$, where x'_i denotes the observed value of X_i, as a function of the parameters θ_i, called the likelihood. It is advocated that the values of L may serve as a measure of our uncertainty or confidence in the corresponding values of the θ's. Accordingly, we should have the greatest confidence in the values, say, $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_l$, for which L is a maximum. $\hat{\theta}_i$ obviously is a function of $x'_1 \ldots x'_n$; it is called the maximum likelihood estimate of θ_i.

As far as I am aware, the idea of the maximum likelihood estimates is due to KARL PEARSON, who applied the principle in 1895 (*see* particularly pp. 262–265), among others to deduce the now familiar formula for estimating the coefficient of correlation. However, he did not insist much on the general applicability of the principle. This was done with great emphasis by R. A. FISHER, who invented the term likelihood, and in a series of papers (FISHER, 1925) stated several important properties of the maximum likelihood estimates, to the general effect that it is improbable that their values will differ very much from those of the parameters estimated. In fact, the maximum likelihood estimates appear to be what could be called the best " almost unbiassed " estimates. Many of FISHER's statements, partly in a modified form, were subsequently proved by HOTELLING (1932), DOOB (1934), and DUGUÉ (1936). An excellent account of the present state of the theory is given by DARMOIS (1936).

In certain cases the unbiassed estimates are identical with those of maximum likelihood; in others we know only the maximum likelihood estimate, and then there is no " competition " between the two principles. But it sometimes happens that both principles may be applied and lead to different results. Such is, for instance, the case when it is known that the variables (4) are all independent and each of them follows the same normal law, so that

$$p(E|\xi, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-\frac{\Sigma(x_i-\xi)^2}{2\sigma^2}}. \quad \ldots\ldots\ldots \quad (13)$$

The maximum likelihood estimate of the variance, σ^2, is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \ldots\ldots\ldots \quad (14)$$

while the unbiassed estimate is, say,

$$\bar{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 , \quad \ldots\ldots\ldots \quad (15)$$

and the question arises which of them to use. Obviously this is a question of principle, and the arguments, like " you must use (15) because the expectation of $\bar{\sigma}^2$ is equal to σ^2 ", do not prove much by themselves. It is perhaps remarkable that some of the authors who, when discussing theory, advocate the use of the maximum likelihood estimate, use in practice the estimate (15).

The formulae (14) and (15) may be used to illustrate the meaning of the expression " almost unbiassed " estimate, used above. Familiar formulae show that the expectation of $\hat{\sigma}^2$ is

$$\mathscr{E}(\hat{\sigma}^2) = \left(1 - \frac{1}{n}\right)\sigma^2 , \quad \ldots\ldots\ldots \quad (16)$$

thus showing a " negative bias," $n^{-1}\sigma^2$. If we increase the number of observations, n, the bias tends to zero, which justifies the terms " almost unbiassed " or " consistent " estimate attached to (14).

(c) *Estimation by Unique Estimate and by Interval*

In the preceding pages we have described briefly three of the several important principles advanced for the calculation of estimates. All of them represent attempts to solve the problem which might be called the problem of a unique estimate of an unknown parameter which reduces to determining a function of the observations, the value of which presumably does not differ very much from that of the estimated parameter.

We shall now call attention to the fact that apart from the problem of a unique estimate, the requirements of practical statistical work brought to the front another problem which we shall call the problem of estimation by interval.

Denote generally by θ the parameter to be estimated and by T its estimate, deduced from some principle or another. Whatever the principle, it is obviously impossible to assume that in any particular case T is exactly equal to θ. Therefore, the practical statistician required some measure of the accuracy of the estimate T. The generally accepted method of describing this accuracy consists in calculating the estimate, say S_T^2, of the variance V_T of T and in writing the result of all the calculations in the form $T \pm S_T$.

Behind this method of presenting the results of estimating θ, there is the idea that the true value of θ will frequently lie between the value of T minus a certain multiple of S_T and T plus perhaps some other multiple of S_T. Therefore, the smaller S_T the more accurate is the estimate T of θ.

If we look through a number of recent statistical publications, we shall find that it is exceedingly rare that the values of unique estimates are given without the $\pm S_T$.

We shall find also that the comments on the values of T are largely dependent on those of S_T. This shows that what the statisticians have really in mind in problems of estimation is not the idea of a unique estimate but that of two estimates having the form, say

$$\underline{\theta} = T - k_1 S_T \quad \text{and} \quad \overline{\theta} = T + k_2 S_T, \quad \ldots\ldots\ldots \quad (17)$$

where k_1 and k_2 are certain constants, indicating the limits between which the true value of θ presumably falls.

In this way the practical work, which is frequently in advance of the theory, brings us to consider the theoretical problem of estimating the parameter θ by means of the interval $(\underline{\theta}, \overline{\theta})$, extending from $\underline{\theta}$ to $\overline{\theta}$. These limits will be called the lower and upper estimates of θ respectively. It is obvious that if the values of k_1 and k_2 in (17) are not specified, then the real nature of the two estimates is not determined.

In what follows, we shall consider in full detail the problem of estimation by interval. We shall show that it can be solved entirely on the ground of the theory of probability as adopted in this paper, without appealing to any new principles or measures of uncertainty in our judgements. In so doing, we shall try to determine the lower and upper estimates, $\underline{\theta}$ and $\overline{\theta}$, which assure the greatest possible accuracy of the result, without assuming that they must necessarily have the commonly adopted form (17).

II—Confidence Intervals

(*a*) *Statement of the Problem*

After these somewhat long preliminaries, we may proceed to the statement of the problem in its full generality.

Consider the variables (4) and assume that the form of their probability law (5) is known, that it involves the parameters $\theta_1, \theta_2, \ldots, \theta_l$, which are constant (not random variables), and that the numerical values of these parameters are unknown. It is desired to estimate one of these parameters, say θ_1. By this I shall mean that it is desired to define two functions $\overline{\theta}(E)$ and $\underline{\theta}(E) \leq \overline{\theta}(E)$, determined and single valued at any point E of the sample space, such that if E′ is the sample point determined by observation, we can (1) calculate the corresponding values of $\underline{\theta}(E')$ and $\overline{\theta}(E')$, and (2) state that the true value of θ_1, say θ_1^0, is contained within the limits

$$\underline{\theta}(E') \leq \theta_1^0 \leq \overline{\theta}(E'), \quad \ldots\ldots\ldots\ldots \quad (18)$$

this statement having some intelligible justification on the ground of the theory of probability.

This point requires to be made more precise. Following the routine of thought established under the influence of the Bayes Theorem, we could ask that, given the sample point E′, the probability of θ_1^0 *falling within the limits* (18) should be large, say, $\alpha = 0{\cdot}99$, etc. If we express this condition by the formula

$$P\{\underline{\theta}(E') < \theta_1^0 < \overline{\theta}(E') | E'\} = \alpha, \quad \ldots\ldots\ldots \quad (19)$$

we see at once that it contradicts the assumption that θ_1^0 is constant. In fact, on this assumption, whatever the fixed point E′ and the values $\underline{\theta}$ (E′) and $\bar{\theta}$ (E′), the only values the probability (19) may possess are zero and unity. For this reason we shall drop the specification of the problem as given by the condition (19).

Returning to the inequalities (18), we notice that while the central part, θ_1^0, is a constant, the extreme parts $\underline{\theta}$ (E′) and $\bar{\theta}$ (E′) are particular values of random variables. In fact, the coordinates of the sample point E are the random variables (4), and if $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are single-valued functions of E, they must be random variables themselves.

Therefore, whenever the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are defined in one way or another, but the sample point E is not yet fixed by observation, we may legitimately discuss the probability of $\underline{\theta}$ (E) and $\bar{\theta}$ (E) fulfilling any given inequality and in particular the inequalities analogous to (18), in which, however, we must drop the dashes specifying a particular fixed sample point E′. We may also try to select $\underline{\theta}$ (E) and $\bar{\theta}$ (E) so that the probability of $\underline{\theta}$ (E) falling short of θ_1^0 and at the same time of $\bar{\theta}$ (E) exceeding θ_1^0, is equal to any number α between zero and unity, fixed in advance. If θ_1^0 denotes the true value of θ_1, then of course this probability must be calculated under the assumption that θ_1^0 *is* the true value of θ_1. Thus we can look for two function $\underline{\theta}$ (E) and $\bar{\theta}$ (E), such that

$$P\{\underline{\theta}\,(E) \leq \theta_1^0 \leq \bar{\theta}\,(E) \mid \theta_1^0\} = \alpha. \quad \ldots\ldots\ldots \quad (20)$$

and require that the equation (20) holds good *whatever* the value θ_1^0 of θ_1 and *whatever* the values of the other parameters $\theta_2, \theta_3, \ldots, \theta_l$, involved in the probability law of the X's may be.

The functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) satisfying the above conditions will be called the lower and the upper confidence limits of θ_1. The value α of the probability (20) will be called the confidence coefficient, and the interval, say δ (E), from $\underline{\theta}$ (E) to $\bar{\theta}$ (E), the confidence interval corresponding to the confidence coefficient α.

It is obvious that the form of the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) must depend upon the probability law $p\,(E|\theta_1 \ldots \theta_l)$.

It will be seen that the solution of the mathematical problem of determining the confidence limits $\underline{\theta}$ (E) and $\bar{\theta}$ (E) provides the solution of the practical problem of estimation by interval. For suppose that the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined so that the equation (20) does hold good whatever the values of all the parameters $\theta_1, \theta_2, \ldots \theta_l$ may be, and α is some fraction close to unity, say $\alpha = 0{\cdot}99$. We can then tell the practical statistician that whenever he is certain that the form of the probability law of the X's is given by the function $p\,(E|\theta_1, \theta_2, \ldots \theta_l)$ which served to determine $\underline{\theta}$ (E) and $\bar{\theta}$ (E), he may estimate θ_1 by making the following three steps : (*a*) he must perform the random experiment and observe the particular values $x_1, x_2, \ldots x_n$ of the X's ; (*b*) he must use these values to calculate the corresponding values of $\underline{\theta}$ (E) and $\bar{\theta}$(E) ; and (*c*) he must state that $\underline{\theta}\,(E) < \theta_1^0 < \bar{\theta}\,(E)$, where θ_1^0 denotes the true value of θ_1. How can this recommendation be justified ?

The justification lies in the character of probabilities as used here, and in the law of great numbers. According to this empirical law, which has been confirmed by numerous experiments, whenever we frequently and independently repeat a random experiment with a constant probability, α, of a certain result, A, then the relative frequency of the occurrence of this result approaches α. Now the three steps (*a*), (*b*), and (*c*) recommended to the practical statistician represent a random experiment which may result in a correct statement concerning the value of θ_1. This result may be denoted by A, and if the calculations leading to the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are correct, the probability of A will be constantly equal to α. In fact, the statement (*c*) concerning the value of θ_1 is only correct when $\underline{\theta}$ (E) falls below θ_1^0 and $\bar{\theta}$ (E), above θ_1^0, and the probability of this is equal to α whenever θ_1^0 is the true value of θ_1. It follows that if the practical statistician applies permanently the rules (*a*), (*b*) and (*c*) for purposes of estimating the value of the parameter θ_1, in the long run he will be correct in about 99 per cent. of all cases.

It is important to notice that for this conclusion to be true, it is not necessary that the problem of estimation should be the same in all the cases. For instance, during a period of time the statistician may deal with a thousand problems of estimation and in each the parameter θ_1 to be estimated and the probability law of the X's may be different. As far as in each case the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are properly calculated and correspond to the same value of α, his steps (*a*), (*b*), and (*c*), though different in details of sampling and arithmetic, will have this in common—the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α.

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α.* Consider now the case when a sample, E′, is already drawn and the calculations have given, say, $\underline{\theta}$ (E′) = 1 and $\bar{\theta}$ (E′) = 2. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α ?

The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for the hypothetical and trivial ones

$$P\{1 \leq \theta_1^0 \leq 2\} = \left\{ \begin{array}{ll} 1 & \text{if} \quad 1 \leq \theta_1^0 \leq 2 \\ 0 & \text{if either } \theta_1^0 < 1 \quad \text{or} \quad 2 < \theta_1^0, \end{array} \right\} \quad . \quad . \quad . \quad (21)$$

which we have decided not to consider.

The theoretical statistician constructing the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E), having the above property (20), may be compared with the organizer of a game of chance in which the gambler has a certain range of possibilities to choose from while, whatever

* This, of course, is subject to restriction that the X's considered *will* follow the probability law assumed.

he actually chooses, the probability of his winning and thus the probability of the bank losing has permanently the same value, $1 - \alpha$.

The choice of the gambler on what to bet, which is beyond the control of the bank, corresponds to the uncontrolled possibilities of θ_1 having this or that value. The case in which the bank wins the game corresponds to the correct statement of the actual value of θ_1. In both cases the frequency of " successes " in a long series of future " games " is approximately known. On the other hand, if the owner of the bank, say, in the case of roulette, knows that in a particular game the ball has stopped at the sector No. 1, this information does not help him in any way to guess how the gamblers have betted. Similarly, once the sample E′ is drawn and the values of $\underline{\theta}$ (E′) $\bar{\theta}$ and (E′) determined, the calculus of probability adopted here is helpless to provide answer to the question of what is the true value of θ_1.

(*b*) *Solution of the Problem of Confidence Intervals*

In order to find the solution of the problem of confidence intervals, let us suppose that it is already solved and that $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are functions determined and single valued in the whole sample space, W, and such that the equality (20) holds good whatever the true values of the parameters $\theta_1, \theta_2, \ldots \theta_l$. It will be convenient to interpret the situation geometrically. For this purpose we shall need to consider the space, G, of $n + 1$ dimensions which we shall call the general space. The points in this space will be determined by $n + 1$ coordinates $x_1, x_2, \ldots x_n, \theta_1$, the first n of which are the particular values of the random variables (4) and thus determine the position of a sample point, E, in the n-dimensional space W, and the last coordinate θ_1 is one of the possible values of the parameter θ_1 in the probability law p (E$|\theta_1 \ldots \theta_l$) which we desire to estimate.

Consequently, if we consider any hyperplane, G (θ_1) in G corresponding to the equation $\theta_1 =$ const., this may be interpreted as an image of the sample space W. We notice also that to any point E in the sample space W there will correspond in G a straight line, say L (E), parallel to the axis Oθ_1. If $x_1', x_2' \ldots x'_n$ are the coordinates of E′, then the line L (E′) will correspond to the equations $x_i = x_i'$ for $i = 1, 2, \ldots n$.

Consider now the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E). On each line L (E), they will determine two points, say B (E) and C (E) with coordinates

$$x_1, x_2, \ldots x_n, \underline{\theta}\ (\mathrm{E}) \quad \ldots\ldots\ldots\ldots \quad (22)$$

and

$$x_1, x_2, \ldots x_n, \bar{\theta}\ (\mathrm{E}) \quad \ldots\ldots\ldots\ldots \quad (23)$$

respectively, where $x_1, x_2, \ldots x_n$ are the coordinates of the sample point E. The interval between B (E) and C (E) will be the image of the confidence interval δ (E) corresponding to the sample point E. If we fix a value of $\theta_1 = \theta_1'$ and a sample point E′, then the hyperplane G (θ_1') may cut or may not cut the confidence interval δ (E′). If G (θ_1') does cut δ (E′), let a (θ_1', E′) denote the point of intersection.

The position is illustrated in fig. 1, in which, however, only three axes of coordinates are drawn, Ox_1, Ox_n, and $O\theta_1$. The line L (E′) is represented by a dotted vertical line and the confidence interval δ (E′) by a continuous section of this line, which is thicker above and thinner below the point a (θ'_1, E′) of its intersection with the hyperplane G (θ_1'). The confidence interval δ (E″) corresponding to another sample point, E″, is not cut by G (θ_1') and is situated entirely above this hyperplane.

Now denote by A (θ_1') the set of all points a (θ'_1, E) in G (θ'_1) in which this hyperplane cuts one or the other of the confidence intervals δ (E), corresponding to any sample point. It is easily seen that the coordinate θ_1 of any point belonging to A (θ'_1) is equal to θ'_1 and that the remaining coordinates $x_1, x_2, \ldots x_n$ satisfy the inequalities

$$\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E). \quad . \; . \quad (24)$$

In many particular problems it is found that the set of points A (θ_1) thus defined is filling up a region. Because of this A (θ'_1) will be called the region of acceptance corresponding to the fixed value of $\theta_1 = \theta'_1$.

It may not seem obvious that the region of acceptance A (θ_1) as defined above must exist (contain points) for any value of θ_1. In fact, it may seem possible that for certain values of θ_1 the hyperplane G (θ_1) may not cut any of the intervals δ (E). It will, however, be seen below that this is impossible.

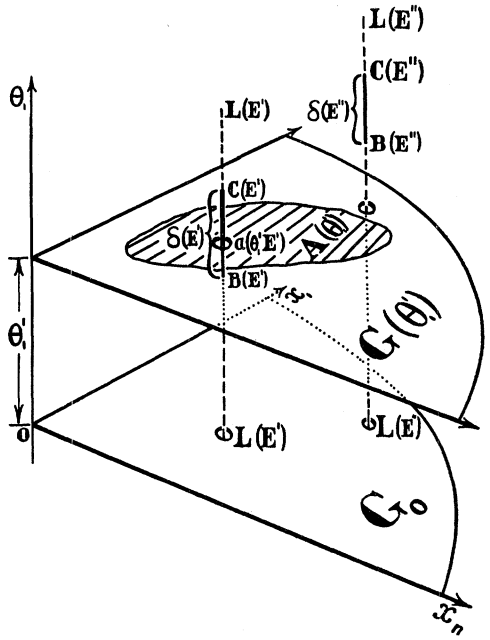

FIG. 1—The general space G.

As mentioned above, the coordinates x_1, $x_2, \ldots x_n$ of any sample point E determine in the space G the straight line L (E) parallel to the axis of θ_1. If this line crosses the hyperplane G (θ_1) in a point belonging to A (θ_1) it will be convenient to say that E falls within A (θ_1).

If for a given sample point E the lower and the upper estimates satisfy the inequalities $\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E)$, where θ'_1 is any value of θ_1, then it will be convenient to describe the situation by saying that the confidence interval δ (E) covers θ'_1. This will be denoted by δ (E) Cθ'_1.

The conception and properties of the regions of acceptance are exceedingly important from the point of view of the theory given below. We shall therefore discuss them in detail proving separately a few propositions, however simple they may seem to be.

Proposition I—Whenever the sample point E falls within the region of acceptance A (θ'_1), corresponding to any fixed value θ'_1 of θ_1, then the corresponding confidence interval δ (E) must cover θ'_1.

Proof—This proposition is a direct consequence of the definition of the region of

acceptance. Suppose it is not true. Then there must be at least one sample point, say E′, which falls within A (θ'_1) and such that either $\underline{\theta}\,(E') \leqq \bar{\theta}\,(E') < \theta'_1$ or $\theta'_1 < \underline{\theta}\,(E') \leqq \bar{\theta}\,(E')$. Comparing these inequalities with (24) which serve as a definition of the region of acceptance A (θ'_1), we see that E′ could not fall within A (θ'_1), which proves the Proposition I.

Proposition II—If a confidence interval $\delta\,(E'')$ corresponding to a sample point E″ covers a value θ'_1 of θ_1, then the sample point E″ must fall within A (θ'_1).

Proof—If $\delta\,(E'')$ covers θ'_1, then it follows that $\underline{\theta}\,(E'') \leqq \theta'_1 \leqq \bar{\theta}\,(E'')$. Comparing these inequalities with (24) defining the region A (θ'_1), we see that E″ must fall within A (θ'_1).

If we agree to denote generally by $\{B \varepsilon A\}$ the words " B belongs to A " or " B is an element of A ", then we may sum up the above two propositions by writing the identity

$$\{E \varepsilon A\,(\theta'_1)\} \equiv \{\delta\,(E)\; C\theta'_1\} \equiv \{\underline{\theta}\,(E) \leqq \theta'_1 \leqq \bar{\theta}\,(E)\}, \quad \ldots \quad (25)$$

meaning that the event consisting in the sample point E falling within the region of acceptance A (θ'_1) is equivalent to the other event which consists in θ'_1 being covered by $\delta\,(E)$.

Corollary I—It follows from the Proposition I and II that whatever may be the true values $\theta'_1, \theta'_2 \ldots \theta'_l$ of the θ's, the probability of any fixed value θ''_1 of θ_1 being covered by $\delta\,(E)$ is identical with the probability of the sample point E falling within A (θ''_1).

$$P\{\delta\,(E)\; C\theta''_1 | \theta'_1, \ldots \theta'_l\} = P\{\underline{\theta}\,(E) \leqq \theta''_1 \leqq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\}$$
$$= P\{E \varepsilon A\,(\theta''_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad (26)$$

Proposition III—If the functions $\underline{\theta}\,(E)$ and $\bar{\theta}\,(E)$ are so determined that whatever may be the true values of $\theta_1, \theta_2, \ldots \theta_l$, the probability, P, of the true value of θ_1 being covered by the interval $\delta\,(E)$ extending from $\underline{\theta}\,(E)$ to $\bar{\theta}\,(E)$ is always equal to a fixed number α, then the region of acceptance A (θ'_1) corresponding to any fixed value θ'_1 of θ_1 must have the property that the probability

$$P\{E \varepsilon A\,(\theta'_1) | \theta'_1, \theta_2, \ldots \theta_l\} = \alpha, \quad \ldots \ldots \quad (27)$$

whatever may be the values of the parameters $\theta_2, \theta_3, \ldots \theta_l$.

Proof—Assume that θ'_1 happens to be the true value of θ_1 and denote generally by θ'_i the true value of θ_i, for $i = 2, 3, \ldots l$. The probability P, as defined in conditions of the Proposition III, may be expressed by means of the formula

$$P = P\{\underline{\theta}\,(E) \leqq \theta'_1 \leqq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad \ldots \ldots \quad (28)$$

Owing to (26), which holds good for any $\theta'_1, \theta'_2, \ldots \theta'_l$, we may write also

$$P = P\{E \varepsilon A\,(\theta'_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad \ldots \ldots \quad (29)$$

If P is permanently equal to α, then $P\{E \varepsilon A(\theta'_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}$ must be also equal to α, whatever $\theta'_1, \theta'_2, \ldots \theta'_l$, which proves the proposition.

Corollary II—It follows from the Proposition III that whatever the value θ'_1 of θ_1, the region of acceptance $A(\theta'_1)$ could not be empty. In fact, if for any value θ'_1 the region $A(\theta_1')$ as defined above did not contain any points at all, then the probability $P\{E \varepsilon A(\theta'_1) | \theta'_1, \ldots \theta'_l\}$ would be zero, which would contradict the Proposition III.

Proposition III describes the fundamental property of any single region of acceptance $A(\theta_1)$ taken separately. We shall now prove three propositions concerning the whole set of the regions $A(\theta_1)$ corresponding to all possible values of θ_1.

Proposition IV—If the functions $\underline{\theta}(E)$ and $\bar{\theta}(E) \geqq \underline{\theta}(E)$ are single valued and determined for any sample point E, then whatever the sample point E', there will exist at least one value of θ_1, say θ'_1, such that the point E' will fall within $A(\theta'_1)$.

Proof—Consider the values of $\underline{\theta}(E)$ and $\bar{\theta}(E)$ corresponding to the point E' and let θ'_1 be any value of θ_1 satisfying the condition $\underline{\theta}(E') \leqq \theta'_1 \leqq \bar{\theta}(E')$. Comparing these inequalities with (24), we see that E' must fall within $A(\theta'_1)$, which proves the proposition.

Proposition V—If a sample point E' falls within the regions of acceptance $A(\theta'_1)$ and $A(\theta''_1)$ corresponding to θ'_1 and $\theta''_1 > \theta'_1$ respectively, then it will also fall within the region of acceptance $A(\theta'''_1)$ corresponding to any θ'''_1 such that $\theta'_1 < \theta'''_1 < \theta''_1$.

Proof—If the sample point E' falls within $A(\theta'_1)$ and $A(\theta''_1)$ then, owing to (24), it follows that

$$\underline{\theta}(E') \leqq \theta'_1 < \theta''_1 \leqq \bar{\theta}(E'). \quad \ldots\ldots\ldots \quad (30)$$

Accordingly, whatever θ'''_1 such that $\theta'_1 < \theta'''_1 < \theta''_1$, it follows that

$$\underline{\theta}(E') < \theta'''_1 < \bar{\theta}(E'), \quad \ldots\ldots\ldots \quad (31)$$

which shows that E' falls within $A(\theta''')$.

Proposition VI—If a sample point E' falls within any of the regions $A(\theta_1)$ for $\theta'_1 < \theta_1 < \theta''_1$ where θ'_1 and θ''_1 are fixed numbers, then it must also fall within $A(\theta'_1)$ and $A(\theta''_1)$.

Proof—Suppose that the proposition is not true and that, for example, E' does not fall within $A(\theta'_1)$. Then it follows that

$$\theta'_1 < \underline{\theta}(E'). \quad \ldots\ldots\ldots\ldots\ldots \quad (32)$$

Let θ'''_1 be a number exceeding θ'_1 but smaller than either $\underline{\theta}(E')$ and θ'', so that $\theta' < \theta''' < \theta''$ and $\theta''' < \underline{\theta}(E')$. It follows from the definition (24) of $A(\theta_1)$ that E' does not fall within $A(\theta''')$, contrary to the assumption that for any θ_1 such that $\theta'_1 < \theta_1 < \theta''_1$ the point E' falls within $A(\theta_1)$. Similarly it is possible to prove that E' must fall within $A(\theta''_1)$.

The four propositions III, IV, V, and VI describe the necessary conditions which must be satisfied by the regions of acceptance A (θ_1), either separately by each of them or collectively, if the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined and single valued in the whole sample space W and if the equation (20) holds good for any value of θ_1; that is to say when they determine the confidence intervals required.

We shall now prove the reciprocal proposition, showing that if it is possible to select on each hyperplane G (θ_1) a region A (θ_1) having the properties as described in the conclusions of the propositions III to VI, then the system of these regions may be used to define the functions $\underline{\theta}$ (E) $\leq \bar{\theta}$ (E) which will be determined and single valued at any sample point E; further, their system will have the property that for any value ${\theta_1}^0$ of θ_1 the equality (20) will hold good, whatever the values of the other parameters $\theta_2, \theta_3, \ldots \theta_l$.

Suppose, therefore, that on each hyperplane G (θ_1) there is defined a region, A′ (θ_1), such that

(i) P $\{E \varepsilon A' (\theta_1) | \theta_1\} = \alpha$, whatever the values of $\theta_2, \theta_3, \ldots \theta_l$.

(ii) Whatever the sample point E, there exists at least one value ${\theta'}_1$ of θ_1 such that E falls within A′ $({\theta'}_1)$.

(iii) If a sample point E falls within A′ $({\theta'}_1)$ and A′ $({\theta''}_1)$ where ${\theta'}_1 < {\theta''}$, then, whatever θ''', such that ${\theta'}_1 < {\theta'''}_1 < {\theta''}_1$, the point E falls within A′ $({\theta'''}_1)$.

(iv) If a sample point E falls within A′ (θ_1) for any θ_1 satisfying the inequalities ${\theta'}_1 < \theta_1 < {\theta''}_1$, then it falls also within A′ $({\theta'}_1)$ and A′ $({\theta''}_1)$.

Denote by $\underline{\theta}'$ (E) the lower and by $\bar{\theta}'$ (E) the upper bound of values of θ_1 for which a fixed sample point E falls within A′ (θ_1).

Proposition VII—If the regions A′ (θ_1) satisfy the conditions (i), (ii), (iii), and (iv), then the functions $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) are the lower and the upper confidence limits of θ_1, *i.e.*, such that

(*a*) they are determined and single valued at any point E and $\underline{\theta}'$ (E) $\leq \bar{\theta}'$ (E),

(*b*) whatever the true value ${\theta_1}^0$ of θ_1, the probability

$$P \{\underline{\theta}' (E) \leq {\theta_1}^0 \leq \bar{\theta}' (E) | {\theta_1}^0\} = \alpha, \quad \ldots\ldots\ldots \quad (33)$$

independently of the values of the other parameters $\theta_2, \theta_3, \ldots \theta_l$.

Proof—The property (*a*) of functions $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) follows directly from the condition (ii) and the definition of $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E). We may notice, however, that $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) are not necessarily finite.

To prove the property (*b*), it will be sufficient to show that whatever ${\theta_1}^0$

$$P \{\underline{\theta}' (E) \leq {\theta_1}^0 \leq \bar{\theta}' (E) | {\theta_1}^0\} = P \{E \varepsilon A' ({\theta_1}^0) | {\theta_1}^0\}, \quad \ldots\ldots \quad (34)$$

and then refer to the condition (i).

For this purpose we notice first that owing to the definition of $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E), whenever E falls within A′ $({\theta_1}^0)$, then it must follow that $\underline{\theta}'$ (E) $\leq {\theta_1}^0 \leq \bar{\theta}'$ (E).

It remains to show that inversely, if for any point E, $\underline{\theta}'(E) \leq \theta_1^0 \leq \bar{\theta}'(E)$, then this point must fall within $A'(\theta_1^0)$.

Suppose for a moment that this is not true and that there is a sample point E′ not falling within $A'(\theta_1^0)$ and such that $\underline{\theta}'(E') \leq \theta_1^0 \leq \bar{\theta}'(E')$.

It is easily seen that in such a case, either $\underline{\theta}'(E') = \theta_1^0$ or $\theta_1^0 = \bar{\theta}'(E')$ or both, if $\underline{\theta}'(E') = \bar{\theta}'(E')$. In fact, if $\underline{\theta}'(E') < \theta_1^0 < \bar{\theta}'(E')$, then $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$, being the lower and the upper bounds of the values of θ_1 for which E′ falls within $A'(\theta_1)$, there would exist two values of θ_1, say θ'_1 and θ''_1, such that E′ is falling within $A'(\theta'_1)$ and $A'(\theta''_1)$ and

$$\underline{\theta}'(E') \leq \theta'_1 < \theta_1^0 < \theta''_1 \leq \bar{\theta}'(E'). \quad \ldots\ldots \quad (35)$$

It would then follow from the condition (iii) that E′ falls within $A'(\theta_1^0)$, contrary to the assumption. Therefore, we cannot assume that $\underline{\theta}'(E') < \bar{\theta}_1^0 < \theta'(E')$.

Now it is easy to see that if

$$\underline{\theta}'(E') = \theta_1^0 = \bar{\theta}'(E') \quad \ldots\ldots\ldots\ldots \quad (36)$$

then E′ must fall within $A'(\theta_1^0)$. In fact, $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$ are respectively the lower and the upper bounds of the values of θ_1 for which E′ falls within $A'(\theta_1)$. If they are both equal to θ_1^0, then θ_1^0 must be the only value of θ_1 for which E′ falls within $A'(\theta_1)$.

It remains to consider only such cases where either $\underline{\theta}'(E') = \theta_1^0 < \bar{\theta}'(E')$ or $\underline{\theta}'(E') < \theta_1^0 = \bar{\theta}'(E')$. In both cases $\underline{\theta}'(E') < \bar{\theta}'(E')$. We notice first that, whatever θ_1, within the limits

$$\underline{\theta}'(E') < \theta_1 < \bar{\theta}'(E') \quad \ldots\ldots\ldots\ldots \quad (37)$$

the sample point E′ must fall within $A'(\theta_1)$. Otherwise either $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$ would not be respectively the lower and the upper bounds of values of θ_1 for which E′ falls within $A'(\theta_1)$, or else the condition (iii) would not be satisfied. Now it follows from (iv) that E′ must fall both within $A'(\theta'_1)$ and $A'(\theta''_1)$ where $\theta'_1 = \underline{\theta}'(E')$ and $\theta''_1 = \bar{\theta}'(E')$ and therefore within $A'(\theta_1^0)$, which completes the proof of the Proposition VII.

Thus the problem of constructing the system of confidence intervals is equivalent to that of selecting on each hyperplane, $G(\theta_1)$, regions $A(\theta_1)$ satisfying the conditions (i)–(iv). Obviously, these regions will have the property of being regions of acceptance.

Before going any further with the theory and discussing the problem of how to choose the most appropriate system of the regions of acceptance, we shall illustrate the results already reached on two examples. These have been selected so as to reduce to a minimum the technical difficulties in carrying out the necessary calculations which might easily conceal the essential points of the theory to be illustrated. It is obvious that under the circumstances the examples could hardly fail to be somewhat artificial. However, at the end of the paper the reader will find examples having direct practical importance.

(c) *Example I*

Consider first the case where the probability law of the random variables considered depends only upon one unknown parameter θ, which it is desired to estimate. Assume further, for simplicity, that the number of random variables, the particular values of which may be given by observation is $n = 2$ and that their elementary probability law $p(x_1, x_2|\theta)$ is known to be

$$\left.\begin{aligned} &p(x_1, x_2|\theta) = \theta^{-2} \quad \text{for} \quad 0 < x_1, x_2 < \theta \\ \text{and} \quad &p(x_1, x_2|\theta) = 0 \quad \text{for any other system of values of } x_1 \text{ and } x_2 \end{aligned}\right\}. \tag{38}$$

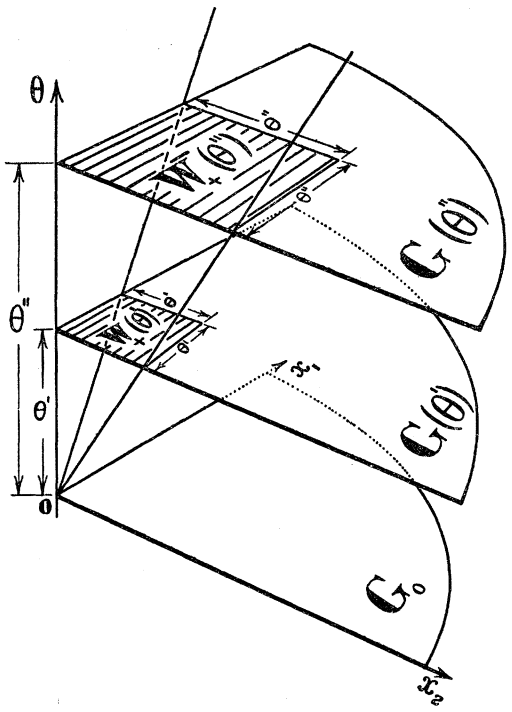

FIG. 2.

The value of θ is unknown and it is desired to construct a system of confidence intervals for its estimation.

The sample space W is now of two dimensions, *i.e.*, a plane. As the coordinates of the sample point must be positive, we may consider that W is limited by the conditions $0 < x_1$ and $0 < x_2$. Denote by $W_+(\theta)$ the part of W in which $p(x_1, x_2|\theta)$ is positive. Obviously $W_+(\theta)$ is a square with its side equal to θ.

Fig. 2 represents the general space G with two planes $G(\theta)$ on which the corresponding squares $W_+(\theta)$ are marked.

According to Proposition VII, the construction of the system of confidence intervals will be completed if we manage to select on each of the planes $G(\theta)$ a region of acceptance $A(\theta)$ satisfying (i)–(iv). Now it is easily seen that it is possible to suggest many systems of regions, some of which will and some of which will not satisfy all these conditions. We shall consider three systems, which will be denoted by S_1, S_2, and S_3, and the particular regions forming these systems by $A_1(\theta)$, $A_2(\theta)$, and $A_3(\theta)$ respectively.

(1) Fix any value of θ and let the region of acceptance $A_1(\theta)$ be defined by the inequalities

$$\beta\theta < x_i < \theta \quad \text{for} \quad i = 1, 2, \ldots\ldots\ldots \tag{39}$$

where β is a positive number less than unity and so selected as to satisfy the condition

$$P\{E_\varepsilon A_1(\theta)|\theta\} = \alpha. \ldots\ldots\ldots\ldots \tag{40}$$

Obviously

$$P\{E_\varepsilon A_1(\theta)|\theta\} = (1 - \beta)^2, \ldots\ldots\ldots \tag{41}$$

and it follows that

$$\beta = 1 - \alpha^{\frac{1}{2}}. \quad \ldots\ldots\ldots\ldots \quad (42)$$

The regions $A_1(\theta)$ defined by (39) will form the system S_1. If β is selected as indicated in (42), they will satisfy the condition (i). Now it is easy to see that they will not satisfy the condition (ii) and that therefore the system S_1 does not present a suitable choice of regions of acceptance which would determine the confidence intervals.

To see this, take any sample point E′ with coordinates x'_1, x'_2, and see whether it is always possible to find a value of $\theta = \theta'$, such that E′ will fall within $A_1(\theta')$. Owing to (39), such a value θ' should satisfy the inequalities

$$\beta\theta' < x'_1, x'_2 < \theta', \quad \ldots\ldots\ldots\ldots \quad (43)$$

or, if l and L denote respectively the smaller and the greater of the numbers x'_1 and x'_2, then

$$L < \theta' < l\beta^{-1}. \quad \ldots\ldots\ldots\ldots \quad (44)$$

This shows that the value θ' such that E′ falls within $A_1(\theta'_1)$ can be found only if $L < l\beta^{-1}$, or $\beta L < l$. Now if $l = x'_1 \leqq x'_2 = L$, then these inequalities lead to the condition $\beta x'_2 < x'_1$. If, on the contrary, $l = x'_2 \leqq x'_1 = L$, then $\beta x'_1 < x'_2$. Accordingly, none of the sample points E″ with coordinates x''_1 and x''_2 such that either

$$0 < x''_2 < \beta x''_1 \quad \text{or} \quad 0 < x''_1 < \beta x''_2 \quad \ldots\ldots \quad (45)$$

will fall within any of the regions $A_1(\theta)$ forming the system S_1, and it follows that they could not serve as regions of acceptance. Fig. 3 (i) illustrates the situation. Here cross-hatched areas correspond to (45).

(2) The second system S_2 of regions $A_2(\theta)$ we shall consider might be suggested by intuition. It follows from the definition of the probability law $p(x_1, x_2|(\theta)$ that x_1 and x_2 are mutually independent and that they vary from zero to θ. Under these circumstances, the mean $\bar{x} = \frac{1}{2}(x_1 + x_2)$ will vary symmetrically about $\frac{1}{2}\theta$ and therefore $2\bar{x} = x_1 + x_2 = T$ could be considered as an estimate of θ itself.

Denote by $A_2(\theta)$ a region in $G(\theta)$ defined by the inequalities

$$\theta - \Delta \leqq x_1 + x_2 \leqq \theta + \Delta, \quad \ldots\ldots\ldots \quad (46)$$

where Δ is so selected as to have $P\{E \varepsilon A_2(\theta)|\theta\} = \alpha$. Simple calculations give

$$P\{E \varepsilon A_2(\theta)|\theta\} = 1 - \left(\frac{\Delta}{\theta}\right)^2 = \alpha, \quad \ldots\ldots\ldots \quad (47)$$

and it follows that $\Delta = \theta(1 - \alpha)^{\frac{1}{2}}$. Substituting this value in (46), we get

$$\theta(1 - (1 - \alpha)^{\frac{1}{2}}) \leqq x_1 + x_2 \leqq \theta(1 + (1 - \alpha)^{\frac{1}{2}}) \quad \ldots\ldots \quad (48)$$

as the final definition of the region $A_2(\theta)$. Fig. 3 (ii) shows the form of the region.

It is easily seen that the system S_2 of regions thus defined satisfies all the conditions (i)–(iv).

For example, in order to check the condition (ii), we may notice that whatever the positive numbers x'_1 and x'_2, the value

$$\theta' = \frac{x'_1 + x'_2}{1 - (1 + \alpha)^{\frac{1}{2}}} \quad \ldots \ldots \ldots \ldots \quad (49)$$

satisfies the inequalities (48) which means that the sample point E′ with coordinates x'_1 and x'_2 falls within $A_2(\theta')$.

The other conditions (iii) and (iv) are checked equally easily. Thus the regions $A_2(\theta)$ may be considered as regions of acceptance. Let us now see how they determine the lower and the upper confidence limits of θ, say $\underline{\theta}_2(E)$ and $\bar{\theta}_2(E)$. According to the definition, $\underline{\theta}_2(E)$ is the lower bound of the values θ' of θ for which the sample point E falls within $A_2(\theta')$. If x_1 and x_2 are the coordinates of E, then it follows from (48) that θ' could not be smaller than, but may be as small as, $(x_1 + x_2)(1 + (1 - \alpha)^{\frac{1}{2}})^{-1}$, which means that

$$\underline{\theta}_2(E) = \frac{x_1 + x_2}{1 + (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \ldots \quad (50)$$

Similarly we get from (48) that θ' may be as large as, but could not exceed, $(x_1 + x_2)(1 - (1 - \alpha)^{\frac{1}{2}})^{-1}$, which shows that

$$\bar{\theta}_2(E) = \frac{x_1 + x_2}{1 - (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \ldots \quad (51)$$

Formerly we used the symbol $\delta(E)$ to denote the confidence interval extending from $\underline{\theta}(E)$ to $\bar{\theta}(E)$. Now we shall use the same symbol to denote the *length* of the confidence interval. We shall have from (50) and (51), say

$$\delta_2(E) = \bar{\theta}_2(E) - \underline{\theta}_2(E) = 2(x_1 + x_2)\frac{\sqrt{1 - \alpha}}{\alpha}. \quad \ldots \quad (52)$$

Now we may use (50) and (51) for estimating θ. If the observations provided the values of x_1 and x_2, say x'_1 and x'_2, we should state that

$$\frac{x'_1 + x'_2}{1 + (1 - \alpha)^{\frac{1}{2}}} \leq \theta \leq \frac{x'_1 + x'_2}{1 - (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \quad (53)$$

Whatever value of α may be fixed in advance, such that $0 < \alpha < 1$, we may be certain that the frequency of the statement in the form (53) being correct will, in the long run, approach α.

The accuracy of estimation corresponding to a fixed value of α may be measured by the lengths of the confidence intervals (52).

(3) The regions $A_3(\theta)$ forming the third set, S_3, will be defined by the inequalities

$$q\theta \leq L < \theta \quad \ldots\ldots\ldots\ldots \quad (54)$$

where L denotes again the larger of the two numbers x_1 and x_2, and q a number between zero and unity to be determined so as to satisfy the condition (i)

$$P\{E \varepsilon A_3(\theta) | \theta\} = P\{q\theta \leq L < \theta | \theta\} = \alpha. \quad \ldots\ldots\ldots \quad (55)$$

Fig. 3 (iii) shows the relationship between $W_+(\theta)$ and $A_3(\theta)$ which lies outside the square adjoining the origin of coordinates with its side equal to $q\theta$.

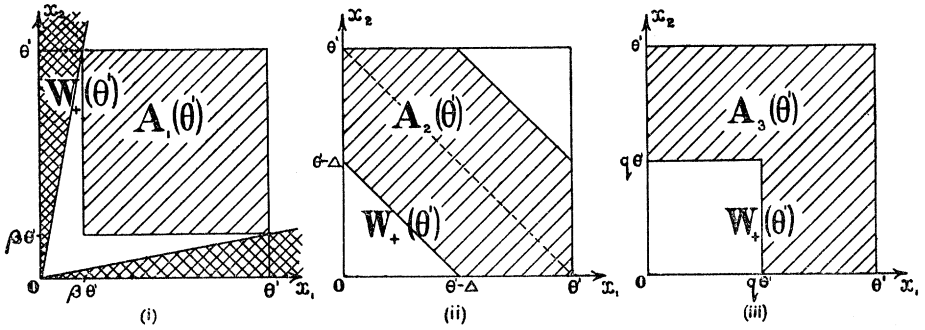

FIG. 3.

It may be useful to deduce the probability law of L for a more general case, when the number n of the x's considered is arbitrary, all of them being independent and following the same probability law

$$\left.\begin{aligned} p(x_i) &= 1/\theta \qquad \text{for } 0 < x_i < \theta \\ p(x_i) &= 0 \qquad \text{elsewhere} \end{aligned}\right\} \quad \ldots\ldots\ldots \quad (56)$$

For this purpose we notice that for any positive constant $L' \leq \theta$

$$P\{L < L' | \theta\} = \prod_{i=1}^{n} \int_0^{L'} p(x_i)\, dx_i = \left(\frac{L'}{\theta}\right)^n \quad \ldots\ldots\ldots \quad (57)$$

Differentiating this expression with regard to L', we may obtain the elementary probability law of L. The probability in the left-hand side of (55) may be obtained directly from (57) and we have, for $n = 2$,

$$P\{q\theta \leq L < \theta | \theta\} = 1 - q^2 = \alpha. \quad \ldots\ldots\ldots \quad (58)$$

Hence

$$q = (1 - \alpha)^{\frac{1}{2}}. \quad \ldots\ldots\ldots\ldots \quad (59)$$

Thus the inequality (54) defining the region $A_3(\theta)$ becomes

$$\theta(1 - \alpha)^{\frac{1}{2}} \leq L < \theta. \quad \ldots\ldots\ldots\ldots \quad (60)$$

It is easily seen that the system S_3 satisfies the conditions (i)–(iv) and therefore may be considered as a system of regions of acceptance defining the lower and the upper confidence limits of θ and hence the confidence intervals. In order to obtain the lower limit, $\underline{\theta}$ (E), fix any sample point E and consider (54). It is easily seen that if L is the larger of the coordinates of E, then the lower bound of the θ's for which E falls within A_3 (θ) is given by

$$\underline{\theta}_3 (E) = L. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (61)$$

On the other hand, it is seen also from (54) that the upper bound of the same θ's is obtained from $q\bar{\theta}$ (E) = L, thus

$$\bar{\theta}_3 (E) = L (1 - \alpha)^{-\frac{1}{2}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (62)$$

It follows that the length of the confidence interval is, say,

$$\delta_3 (E) = \frac{1 - (1 - \alpha)^{\frac{1}{2}}}{(1 - \alpha)^{\frac{1}{2}}} L. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (63)$$

The formulae (61) and (62) could be used to estimate θ, and in applying them we shall be correct, in the long run, in 100α per cent. of all cases.

It is interesting to compare the two systems of confidence intervals (50) and (51), (61) and (62). For this purpose let us choose $\alpha = \frac{3}{4}$. The statements concerning the value of θ using the two confidence intervals will be

$$\tfrac{4}{3}\bar{x} \leq \theta \leq 4\bar{x}, \quad \delta_2 (E) = \tfrac{8}{3}\bar{x}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (64)$$

and

$$L \leq \theta \leq 2L, \quad \delta_3 (E) = L, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (65)$$

where $\bar{x}$ is the arithmetic mean of x_1 and x_2. Assume that in two different cases, A and B, the observations gave $x'_1 = x'_2 = 1$ and $x''_1 = 0{\cdot}1$, $x''_2 = 1{\cdot}9$ respectively. Then using (64) we shall get, in both cases,

$$\tfrac{4}{3} \leq \theta \leq 4, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (66)$$

while using (65)

$$1 \leq \theta \leq 2 \quad \text{and} \quad 1{\cdot}9 \leq \theta \leq 3{\cdot}8 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (67)$$

in cases A and B respectively.

The two pairs of inequalities do not agree and a superficial examination may lead to the conclusion that there is some contradiction in the theory.

It is perhaps not so bad with the sample A, for which the two confidence intervals (66) and (67) partly overlap but do not cover each other. But in the case of the sample B the interval (67) is entirely included within (66). Are these intervals equally reliable ?

Before this question could be answered, it must be made more precise. What is exactly meant by the words " equally reliable ", and do they refer to the numerically defined intervals, viz., (4/3, 4) and (1·9, 3·8), or to the whole systems of intervals as given by (64) and (65) ?

The theory of confidence intervals as explained in preceding pages does give reasons for considering the systems (64) and (65) as " equally reliable ". By this is meant that (1) if a random experiment determining the values of x_1 and x_2 is performed many times and (2) if the random variables x_1 and x_2 follow the probability law (38) where the value of $\theta > 0$ in each experiment may be the same or different—without any limitation whatsoever—then the frequency of cases where the intervals (64) and (65) calculated for each experiment would actually cover the true value of θ will be, in the long run, the same, namely, $\alpha = 3/4$.

On the other hand, if the words " equally reliable " in the above question refer to the numerical intervals (4/3, 4) and (1·9, 3·8), then the theory of confidence intervals does not give any reasons for judging them equally reliable or not.

It may be useful to illustrate the above statements with a simple sampling experiment which the reader may wish to perform.

Imagine that in a period of time the statistician is faced 400 times with the problem of estimating θ. The true value of θ may be in all those 400 cases the same, or it may vary from case to case in an absolutely arbitrary manner. Assume, for instance, that in a set of 400 random experiments the distribution of θ is as set up in the following table (or any other) :

True θ	Frequency
1	155
2	37
10	8
20	10
30	190

Next take TIPPETT's random sample tables (1927) and consider each of the numbers composed of four digits as a decimal fraction. Write down from the table 400 couples of figures. The figures of the first 155 couples consider as particular values of x_1 and x_2 determined by 155 experiments with true $\theta = 1$. The figures in the next 37 couples multiply by 2 and consider the products as forming the results of 37 further experiments where $\theta = 2$. The figures in the next 8 couples should be multiplied by 10, those in the next 10 couples by 20, and finally those in the remaining 190 couples by 30.

Substitute the obtained results in formulae (64) and (65) and see in each case whether the calculated interval covers the true value of θ, *i.e.*, 1, 2, 10, 20, or 30, whichever the case may be. It will be seen that the relative frequency of cases where the confidence intervals either calculated from (64) or from (65) will actually cover the true θ will be approximately equal to $\alpha = 0{\cdot}75$. Of course, there will be no perfect agreement with this figure, but it would be extremely surprising if the observed frequency fell outside the limits of 0·69 and 0·81. This result is entirely independent of the distribution of true θ's, and the above table may be altered as desired, without any limitation.

If there is little to choose between the two systems of confidence intervals (50) and

(51), and (62) and (63) from the point of view of probability of correct statements, there are other aspects which easily determine the choice. In problems of estimation by interval, it is natural to try to get as narrow confidence intervals as possible. Comparing again (66) and (67), we find that the latter interval is considerably shorter than the former. It is easy to see that this is a general rule. In fact, whatever the mean, $\bar{x}$, if both x_1 and x_2 are necessarily positive, then

$$\bar{x} \leqq L < 2\bar{x}, \quad \ldots\ldots\ldots\ldots\ldots\ldots (68)$$

and it follows from (64) and (65) that

$$(3/8)\,\delta_2\,(E) < \delta_3\,(E) < (3/4)\,\delta_2\,(E), \quad \ldots\ldots\ldots (69)$$

showing that the length of the confidence interval determined by (62) and (63) is always less than 3/4 of that determined by (50) and (51). It is obvious, therefore, that the confidence intervals defined by (62) and (63) compared to the other system have definite advantages. These advantages, however, are independent of the conception of probability.

Using again the analogy with the games of chance, we may say that while the rules of the two kinds of game, as described by the two pairs of inequalities (50) and (51), (62) and (63), assure the same probability of winning, the sums which could be won in each case are different, and this is the reason why we prefer the " game " (62) and (63).*

(*d*) *Example II*

Let us now consider an example in which the probability law of the random variables considered depends upon two parameters θ_1 and θ_2, our problem being to estimate the value of θ_1. In order to remove all technical difficulties which might screen the essential points of the theory, we shall again consider a simple case where the number of the random variables is $n = 2$. Suppose that it is known for certain that

$$\left.\begin{aligned} &p\,(x_1, x_2|\theta_1, \theta_2) = \frac{2}{\theta_1^{\,2}} - \theta_1\,\theta_2 + 3\,\theta_2\,x_1 \text{ for } 0 < x_1, x_2 \text{ and } x_1 + x_2 \leqq \theta_1 \\ &p\,(x_1, x_2|\theta_1, \theta_2) = 0 \text{ for any other system of values of the } x\text{'s.} \end{aligned}\right\}. \quad (70)$$

As to the parameters θ_1 and θ_2, it is known only that $\theta_1 > 0$ and $-1 < \theta_1^{\,3}\theta_2 \leqq 2$. The sample space W is limited to the first quadrant of the plane of the x's, and its positive part, $W_+(\theta_1)$, corresponding to any fixed value of θ_1, is formed by a triangle as suggested in fig. 4.

In order to see at once the difficulties introduced by the fact that the probability law (70) depends upon *two* parameters, while we are interested in one only, let us try to solve the problem of confidence intervals by a guess. In Example I, the more

* This point will be discussed later. *See* pp. 370 *et seq.*

satisfactory confidence intervals were determined by regions of acceptance belonging to S_3, having their internal boundary similar to that of the external, the latter being also the external boundary of $W_+(\theta)$.

As the conditions of the problem in Example II present many features similar to those in Example I, let us try to use as regions of acceptance the regions $A_1(\theta_1)$, constructed in the same manner as the more successful regions of acceptance in Example I.

The region $A_1(\theta_1)$ will be limited by the axes of coordinates, by the straight line $x_1 + x_2 = \theta_1$ and by a parallel to that line, corresponding to the equation $x_1 + x_2 = a\,\theta_1$, where $a < 1$ will be a constant which we shall try to determine so as to satisfy the condition (i).

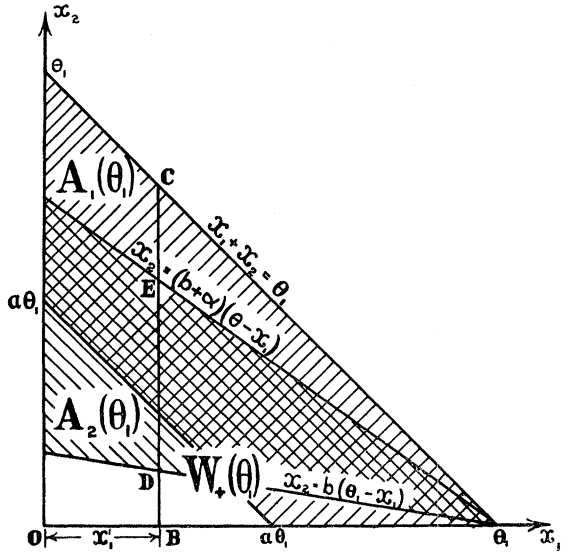

FIG. 4—B C represents $W_+(x'_1)$; D E represents $w(x'_1)$

We have

$$P\{E\varepsilon A_1(\theta_1)|\theta_1, \theta_2\} = P\{a\theta_1 \leq x_1 + x_2 \leq \theta_1|\theta_1, \theta_2\}$$

$$= 1 - \int_0^{a\theta_1} dx_1 \int_0^{a\theta_1 - x_1} p(x_1, x_2|\theta_1, \theta_2)\, dx_2$$

$$= 1 - a^2 + \tfrac{1}{2}a^2(1 - a)\,\theta_1{}^3\theta_2 \quad \ldots\ldots \quad (71)$$

Now it is easy to see that the regions $A_1(\theta_1)$ cannot be used as regions of acceptance.

In fact, it follows from the proposition III that the regions $A_1(\theta_1)$ could only be used as regions of acceptance if, for any fixed value of $\theta_1 = \theta'_1$, the probability $P\{E\varepsilon A_1(\theta'_1)|\theta'_1, \theta_2\}$ were equal to α *irrespective of what is the true value of* θ_2. Looking at the last line of (71), we see that if a and $\theta_1 = \theta'_1$ are fixed, the probability $P\{E\varepsilon A_1(\theta'_1)|\theta'_1, \theta_2\}$ still depends on θ_2 and that, according to the value of this parameter, it may be smaller or larger than the prescribed α.

We see, therefore, that in cases where the probability law of the x's depends upon some more parameters, say $\theta_2, \theta_3, \ldots \theta_l$ besides θ_1, which it is desired to estimate,

the choice of the regions of acceptance must be limited to those, A (θ_1), for which the value of the probability $P\{E \varepsilon A(\theta_1) | \theta_1, \theta_2, \ldots \theta_l\} = \alpha$ and is independent of the values of the parameters $\theta_2 \ldots . \theta_l$.

Regions of this type which have been considered elsewhere (NEYMAN and PEARSON, 1933) are called similar to the sample space with regard to the parameters θ_2, $\theta_3, \ldots \theta_l$, and of size α. If certain limiting conditions are satisfied by the elementary probability law of the X's, it is known also how to construct the most general similar region. Therefore, under these conditions, we are able to select the regions of acceptance, not only satisfying the condition (i) but also some other conditions concerning the relative width of the confidence intervals which will be discussed below.

The conditions under which we are able to construct the most general region similar to the sample space with regard to the parameter θ_2 are not satisfied by the probability law (70). Therefore, we are not able to construct any region similar to W with regard to θ_2. However, a few theoretical remarks which follow allow the construction of a rather broad family, say F, of such regions. It is just possible that an advance of our knowledge on the subject will show that the only regions similar to W with regard to θ_2 are those belonging to F.

(e) *Family of Similar Regions Based on a Sufficient System of Statistics*

Denote by $p(E|\theta_1, \theta_2, \ldots \theta_l)$ the probability law of random variables $X_1, X_2, \ldots X_n$ depending on l parameters $\theta_1, \theta_2, \ldots \theta_l$, by $W(T_1, T_2, \ldots T_s)$, or W (T) for short, the locus of points in the sample space W where some statistics* $T_1, T_2, \ldots T_s$ have certain constant values and finally by $w(T_1, T_2, \ldots T_s)$, or w (T), a part of W (T) which may be defined in one way or another. We shall assume that the T's possess continuous partial derivatives with regard to the X's. We may now prove the following proposition.

Proposition VIII—If the statistics $T_1, T_2, \ldots T_s$ form a sufficient set with regard to the parameters $\theta_2, \theta_3, \ldots \theta_l$, then the probability of the sample point E falling within w (T) calculated under the assumption that it has fallen within W (T) or

$$P\{E \varepsilon w(T) | E \varepsilon W(T)\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (72)$$

is independent of $\theta_2, \theta_3, \ldots \theta_l$ and is a function of θ_1 only.

In proving this proposition, we shall start by expressing its conditions analytically. The condition that the statistics $T_1, T_2, \ldots T_s$ form a sufficient system with regard to $\theta_2, \theta_3, \ldots \theta_l$ is equivalent to (i) that $T_1, T_2, \ldots T_s$ are algebraically independent and (ii) that the elementary probability law of the X's can be presented in the form of the product

$$p(E|\theta_1, \theta_2, \ldots \theta_l) \equiv p(T_1, T_2, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l) f(E|\theta_1), \quad . \quad . \quad (73)$$

* For the definitions of the terms used in this section, *see* NEYMAN and PEARSON (1936, *b*).

where $p\,(T_1, T_2, \ldots T_s \mid \theta_1, \theta_2, \ldots \theta_l)$ means the elementary probability law of the T's and $f\,(E|\theta_1)$ is a function of the x's and possibly of θ_1, but quite independent of $\theta_2, \theta_3, \ldots \theta_l$.* The word " equivalent " means that whenever $T_1, \ldots T_s$ form a sufficient set then both (i) and (ii) must hold good and that, inversely, whenever (i) and (ii) are true, then the statistics $T_1 \ldots T_s$ must form a sufficient set.

Introduce a new system of n-variables $T_1, T_2, \ldots T_s, t_{s+1}, \ldots t_n$, including the statistics T_i, which form the sufficient set, and transforming the original space W of the x's into another n-dimensional space W'. As the T's are algebraically independent, it is always possible to arrange so as to have a one to one correspondence between W and W', except perhaps for a set of points of measure zero. Denoting by E' the point in W' and using (73), we may write the probability law of the new variables in the form

$$p\,(E'|\theta_1, \theta_2, \ldots \theta_l) = p\,(T_1, T_2, \ldots T_s|\theta_1, \ldots \theta_l) f_1\,(E'|\theta_1), \quad . \quad (74)$$

where again $f_1\,(E'|\theta_1)$ does not depend upon $\theta_2, \theta_3 \ldots \theta_l$. Dividing both sides of (74) by $p\,(T_1, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l)$, we shall obtain the relative probability law of $t_{s+1}, t_{s+2}, \ldots t_n$, given $T_1, T_2, \ldots T_s$,

$$p\,(t_{s+1}, t_{s+2}, \ldots t_n|\theta_1, \ldots \theta_l, T_1, \ldots T_s) = f_1\,(E'|\theta_1). \quad . \quad . \quad . \quad (75)$$

Now (72) represents the probability of E falling within w (T), calculated on the assumption that it fell on the hypersurface W (T). The image of W (T) in W' will be a prime, say W'(T), defined by $T_i = \text{const.}$, $i = 1, 2, \ldots s$, and the image of w (T) a part of W' (T), which we shall denote by w' (T). The position of the point E' on W' (T) corresponding to any fixed system of values of $T_1, T_2, \ldots T_s$ is determined by the coordinates $t_{s+1}, t_{s+2}, \ldots t_n$, and it follows that the probability in (72) is equal to the integral of (75) with regard to $t_{s+1}, t_{s+2}, \ldots t_n$ extending over the region w' (T).

As (75) is independent of $\theta_2, \theta_3, \ldots \theta_l$, so must be its integral taken over w' (T),

$$\begin{aligned} P\,\{E \varepsilon w\,(T)|E \varepsilon W\,(T)\} &= P\,\{E' \varepsilon w'\,(T)|E' \varepsilon W'\,(T)\} \\ &= \int \ldots \int_{w'(T)} p\,(t_{s+1}, \ldots t_n|T_1, T_2, \ldots T_s)\, dt_{s+1} \ldots dt_n \\ &= \int \ldots \int_{w'(T)} f_1\,(E'|\theta_1)\, dt_{s+1}\, dt_{s+2} \ldots dt_n \quad . \quad . \quad . \quad . \quad (76) \end{aligned}$$

This completes the proof of the proposition VIII. We may remark that for any fixed value of θ_1 and a fixed system of $T_1, T_2, \ldots T_s$ for which $p\,(T_1, \ldots T_s) > 0$ the region w (T) may be so selected as to ascribe to (76) any value between zero and unity which may be given in advance. It is also obvious that this could be done in an infinity of ways.

* This proposition has been stated without proof by NEYMAN and PEARSON (1936, *b*), p. 121. It may be easily proved following the lines indicated by NEYMAN (1935, *a*).

Proposition IX—If $T_1, T_2, \ldots T_s$ form a sufficient set of statistics with regard to $\theta_2, \theta_3, \ldots \theta_l$ and if for any system of values of the T's the region w (T) is so selected that, for a fixed value of $\theta_1 = \theta'_1$,

$$P\{E\varepsilon w(T)|E\varepsilon W(T)\} = \alpha, \quad \ldots\ldots\ldots \quad (77)$$

where $0 < \alpha < 1$, then, for that value $\theta_1 = \theta'_1$ the n-dimensional region w which would be obtained by combining together the regions w (T) corresponding to all possible systems of values of $T_1, T_2, \ldots T_s$, will be similar to the sample space W with regard to $\theta_2, \theta_3, \ldots \theta_l$ and will have its size equal to α, so that

$$P\{E\varepsilon w|\theta'_1\} = \alpha, \quad \ldots\ldots\ldots\ldots \quad (78)$$

whatever the values of $\theta_2, \theta_3, \ldots \theta_l$.

In order to prove Proposition IX, denote by w' the image of w in W'. Obviously w' will be a combination of the regions w' (T) and also

$$P\{E\varepsilon w|\theta'_1\} = P\{E'\varepsilon w'|\theta'_1\}, \quad \ldots\ldots\ldots\ldots \quad (79)$$

and therefore

$$P\{E\varepsilon w|\theta'_1\} = \int \ldots \int_{w'} p(E'|\theta'_1, \theta_2, \ldots \theta_l)\, dT_1\, dT_2 \ldots dt_n. \quad (80)$$

Using (74) and denoting by W'' the set of all possible systems of values of $T_1, T_2, \ldots T_s$, we obtain further

$$P\{E\varepsilon w|\theta'_1\} = \int \ldots \int_{W''} \Big\{ p(T_1, T_2, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l) \int \ldots \int_{w'(T)} f_1(E'|\theta'_1)\, dt_{s+1} \ldots dt_n \Big\}\, dT_1 \ldots dT_s . \quad (81)$$

Owing to (77), this equation reduces to

$$P\{E\varepsilon w|\theta'_1\} = \alpha \int \ldots \int_{W''} p(T_1, \ldots T_s|\theta'_1, \theta_2, \ldots \theta_l)\, dT_1 \ldots dT_s = \alpha, \quad (82)$$

since the integral of $p(T_1, \ldots T_s|\theta'_1, \ldots \theta_l)$, taken over the set W'' of all possible systems of values of the T's, must be equal to unity, whatever the values of $\theta_1, \theta_2, \ldots \theta_l$. This proves the Proposition IX.

It follows that, whenever a system of statistics $T_1, T_2, \ldots T_s$ sufficient with regard to the parameters $\theta_2, \ldots \theta_l$ exists, we may construct an infinity of regions w, all of which will be similar to the sample space W and will have the same size α. To do so it is sufficient

(*a*) To select on any hypersurface W (T) a region w (T) satisfying the condition (77). Owing to Proposition VIII, this *is* always possible and in an infinity of ways.

(*b*) To combine all the regions w (T) corresponding to all possible systems of values of the T's.

The family of the regions similar to the sample space with regard to $\theta_2, \ldots \theta_l$ which may be thus obtained may be called the family based on the sufficient system of statistics $T_1, T_2, \ldots T_s$. It is possible that in certain cases similar regions will exist which do not enter into such families based on sufficient systems of statistics.

We may now go back to our Example II and see how the problem of confidence intervals could be solved.

(*f*) *Example IIa.*

Turning back to the probability law of x_1 and x_2 as defined in (70), it is easy to see that x_1 is a specific sufficient statistic with regard to θ_2. As a specific sufficient statistic with regard to one parameter is a particular case of a sufficient system of statistics, this fact, together with the Proposition IX, could be used in order to construct regions similar with regard to θ_2, which we require to serve us as regions of acceptance.

In order to see that x_1 is a specific sufficient statistic with regard to θ_2, let us calculate its elementary probability law. Integrating (70) with regard to x_2 between limits zero and $\theta_1 - x_1$, we easily obtain

$$\left.\begin{aligned} &p(x_1) = p(x_1, x_2|\theta_1, \theta_2)(\theta_1 - x_1) \quad \text{for} \quad 0 < x_1 \leqq \theta_1, \\ &p(x_1) = 0 \quad \text{for any other value of } x_1. \end{aligned}\right\} \quad \ldots \quad (83)$$

It is seen that $p(x_1)$ depends both on θ_1 and θ_2 and therefore we shall denote it by $p(x_1|\theta_1\theta_2)$. Now we can write

$$p(x_1, x_2|\theta_1, \theta_2) = p(x_1|\theta_1, \theta_2) f(\mathrm{E}|\theta_1), \quad \ldots \quad (84)$$

with $f(\mathrm{E}|\theta_1)$ defined as follows. For $0 < x_1, x_2$ and $x_1 + x_2 \leqq \theta_1$

$$f(\mathrm{E}|\theta_1) = (\theta_1 - x_1)^{-1}, \quad \ldots \quad (85)$$

and at any other point $f(\mathrm{E}|\theta_1) = 0$. As $f(\mathrm{E}|\theta_1)$ is independent of θ_2, it follows that x_1 is a specific sufficient statistic of θ_2.

Using Proposition IX, we may now construct regions which, for a fixed value of θ_1, will be similar to W with regard to θ_2. For this purpose we have to fix $\theta_1 = \theta'_1$ (say) and also the value of the sufficient statistic $x_1 = x'_1$. Next we consider the locus W (x'_1) where $x = x'$ and select any part of it $w(x')$ satisfying (77).

The combination of $w(x')$ corresponding to all values of x' between limits $0 < x' \leqq \theta'_1$ will give us a region similar to the sample space with regard to θ_2.

Now W (x'_1) is a straight line parallel to the axis Ox_2. In order to select its part $w(x')$, which may be represented by an interval, satisfying (77), we require the relative probability law of x_2, given x_1. Using the familiar relation

$$p(x_1, x_2) = p(x_1)\, p(x_2|x_1), \quad \ldots \quad (86)$$

and comparing it with (84) and (85), we find that for $0 < x_1 \leqq \theta_1$

$$\left.\begin{aligned} p\,(x_2|\theta_1, x_1) &= (\theta_1 - x_1)^{-1} \quad \text{for} \quad 0 < x_2 \leqq \theta_1 - x_1 \\ p\,(x_2|\theta_1, x_1) &= 0 \quad \text{for other values of } x_2. \end{aligned}\right\} \quad . \; . \quad (87)$$

It follows that the relative probability law of x_2, given x_1, is positive and constant for $0 < x_2 \leqq \theta_1 - x_1$ and is zero elsewhere on the line $W(x_1)$. Therefore the condition (77) concerning the interval $w(x'_1)$* to be one of the elements of the similar region w reduces to the requirement that the length of $w(x')$ should be in a constant proportion α to the length of the interval, say $W_+(x'_1)$, on $W(x'_1)$, where $p\,(x_2|\theta_1, x'_1)$ is positive.

We see that a number of regions similar to the sample space with regard to θ_2 could be obtained as follows. (*a*) Fix a value of $x = x' < \theta_1$ and select on the line $W_+(x'_1)$ corresponding to

$$x_1 = x'_1 \quad \text{and} \quad 0 < x_2 \leqq \theta_1 - x'_1, \quad . \; . \; . \; . \; . \; . \; . \quad (88)$$

any interval $w(x'_1)$, the length of which is equal to $\alpha\,(\theta_1 - x'_1)$. (*b*) Combine all such intervals together to form w.

We shall select as the regions of acceptance, $A_2(\theta_1)$, the regions constructed as described in (*a*) and (*b*) with an additional limitation, that the intervals $w(x_1)$ corresponding to different values of x_1 should be similarly situated on $W_+(x_1)$. Thus, for any $0 < x_1 < \theta_1$ we shall define the interval $w(x_1)$ by the inequalities

$$b\,(\theta_1 - x_1) < x_2 \leqq (b + \alpha)\,(\theta_1 - x_1), \quad . \; . \; . \; . \; . \; . \quad (89)$$

where b is any positive number not exceeding $1 - \alpha$. Combining all such intervals, which obviously satisfy (*a*), we shall obtain the region $A_2(\theta_1)$ which we shall use as a region of acceptance in estimating θ_1. As shown in fig. 4, the region $A_2(\theta_1)$ is limited by the axis Ox_2, and by two straight lines $x_2 = b\,(\theta_1 - x_1)$ and $x_2 = (b + \alpha)\,(\theta_1 - x_1)$. It is easy to check that $P\{E \varepsilon A_2(\theta_1)|\theta_1\} = \alpha$ whatever the value of θ_2, so that the condition (i) required for $A_2(\theta_1)$ to be a region of acceptance is satisfied. It is easily seen that the remaining conditions (ii)–(v) are also satisfied.

Now we may determine the confidence intervals for θ_1 resulting from the regions of acceptance $A_2(\theta_1)$. If x'_1 and x'_2 are the coordinates of any sample point E′ determined by observation, we see from (89) that the lower bound of values θ'_1 of θ_1 for which $E' \varepsilon A_2(\theta'_1)$ is

$$\underline{\theta}_1(E') = x'_1 + \frac{x'_2}{b + \alpha} \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (90)$$

* It is obvious that it is not necessary that $w(x')$ should be one single interval on $W_+(x')$. It could be formed by several such intervals subject to the condition that the sum of their lengths is equal to $\alpha\,(\theta_1 - x')$, *etc.*

The upper bound of θ'_1 is found from the same inequalities (89), namely,

$$\bar{\theta}_1 (E') = x'_1 + \frac{x'_2}{b}. \quad \ldots\ldots\ldots\ldots \quad (91)$$

These are two estimates of θ_1 determining the confidence interval δ (E'). The length of this interval for any given sample point, E, is

$$\delta (E) = \frac{\alpha x_2}{b (b + \alpha)}, \quad \ldots\ldots\ldots\ldots \quad (92)$$

and depends upon the value of b chosen. The larger b, the smaller δ (E) and therefore the more accurate estimation of θ_1. The confidence intervals giving the greatest accuracy correspond to $b = 1 - \alpha$.

We see again that after having assured that the probability of our being correct in statements concerning the estimated parameter is equal to α, we can proceed further and satisfy some requirements concerning the accuracy of these statements as measured by the length of the confidence intervals.

The above two examples are simple not only because they do not present any technical difficulties in calculating probability laws, etc., but also because the choice between the systems of confidence intervals suggested is easy, *e.g.*, if we use alternatively $b' = 1 - \alpha$ and $b'' < 1 - \alpha$, all the confidence intervals as determined by (90) and (91) corresponding to b' will be shorter than those corresponding to b''. There is therefore no doubt as to what value of b should be chosen.

This, however, is not always the case, and in general there are two or more systems of confidence intervals possible corresponding to the same confidence coefficient α, such that for certain sample points, E', the intervals in one system are shorter than those in the other, while for some other sample points, E'', the reverse is true.

This point is of some importance and I advise the reader, as a useful exercise, to consider a system of regions of acceptance, $A_3 (\theta_1)$, defined as follows :

(1) for $0 < x_1 \leq 1/2\ \theta_1$, $A_3 (\theta_1)$ contains all points in which

$$(1 - \alpha) (\theta_1 - x_1) \leq x_2 \leq \theta_1 - x_1, \quad \ldots\ldots\ldots \quad (93)$$

(2) for $1/2\ \theta_1 < x_1 < \theta_1$, $A_3 (\theta_1)$ contains all points in which

$$0 < x_2 < \alpha (\theta_1 - x_1). \quad \ldots\ldots\ldots\ldots \quad (94)$$

It is easy to see that the regions $A_3 (\theta_1)$ thus defined may serve as regions of acceptance. The reader will also easily find that for all sample points of the line $x_2 = \alpha x_1$ the confidence intervals as defined by regions $A_3 (\theta_1)$ will be shorter than those defined by (90) and (91) with $b = 1 - \alpha$. On the contrary, the confidence intervals for all sample points lying on the line $x_2 = q x_1$ with

$$0 < q < \frac{\alpha (1 - \alpha)}{1 - \alpha + \alpha^2}, \quad \ldots\ldots\ldots\ldots \quad (95)$$

will be greater than those defined by (90) and (91). The position is illustrated in fig. 5. Here it is not so clear which of the two systems of confidence intervals to choose. The analysis of the situation is given in the next section.

III—Accuracy of Confidence Intervals

(*a*) *Shortest Systems of Confidence Intervals*

If there are possible the systems of confidence intervals, say C_1 and C_2, such that for some sample points the intervals in C_1 are shorter than those in C_2, while for some other sample points the reverse is true, the choice between C_1 and C_2 may be based on the relative frequency or on the probability of having an interval of a given length. If using C_1 we have short confidence intervals more frequently than when using C_2, then the system C_1 will be probably considered as more satisfactory.

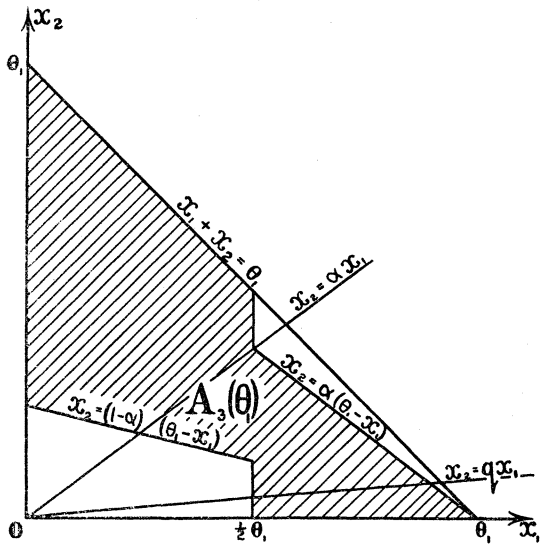

Fig. 5—Shaded area represents $A_3(\theta_1)$

The above statement may appeal to intuition, but it is obviously too vague to be used in practice.

Consider the general problem when the number n of the variables X which we may observe is arbitrary and the probability law of the X's, $p(E|\theta_1, \ldots \theta_l)$ depends on l parameters $\theta_1, \ldots \theta_l$, the first of which, θ_1, we desire to estimate. Denote by θ_1^0 the unknown true value and by θ'_1 any other value of the estimated parameter. Denote further by $\delta_i(E)$ the confidence interval for θ_1 corresponding to the sample point E and belonging to a particular system C_i, $(i = 1, 2 \ldots)$ of the confidence intervals established at a fixed confidence coefficient α. Thus we assume that, as in the above examples, we have several systems of confidence intervals $C_1, C_2, \ldots$ If all of them correspond to the same confidence coefficient α, then all of them satisfy the condition

$$P\{\delta(E)\, C\theta_1^0 | \theta_1^0\} \equiv \alpha, \quad \ldots\ldots\ldots \quad (96)$$

stating that, whatever θ_1^0 and whatever the values of other parameters $\theta_2, \ldots \theta_l$, the probability that the interval should cover the true value θ_1^0, is equal to α.

This is the common property of the systems of confidence intervals considered.

Now it is obvious that whilst it is desirable that the true value of $\theta_1 = \theta_1^0$ should be covered by the confidence interval $\delta(E)$ determined by an observed sample point E, it is not so with any other value of $\theta_1 = \theta'_1 \neq \theta_1^0$. In fact, the presence of the value $\theta'_1 \neq \theta_1^0$ within an interval $\delta(E)$ containing θ_1^0 is unnecessary and may be

interpreted as an indication that this interval is " too broad ". It is clearly impossible to avoid altogether covering the values of θ_1 which are not true. But we may try to diminish the frequency of δ (E) covering any value $\theta'_1 \neq \theta_1^0$ to a minimum. This leads us to the following definition of the shortest system of confidence intervals.

If a system, C_0, of confidence intervals δ_0 (E) has the property that whatever any other system C of intervals δ (E) corresponding to the same confidence coefficient α, whatever the true value of $\theta_1 = \theta_1^0$ and whatever any other value $\theta'_1 \neq \theta_1^0$

$$P\{\delta_0 (E)\, C\theta'_1 | \theta_1^0\} \leq P\{\delta (E)\, C\theta'_1 | \theta_1^0\}, \quad \ldots\ldots \quad (97)$$

then the system C_0 will be called the shortest system of confidence intervals.

The justification of this terminology is clear. When using C_0, the true value of $\theta_1 = \theta_1^0$ will be covered with the prescribed frequency α and any other value $\theta'_1 \neq \theta_1^0$, with a frequency not exceeding that corresponding to any other system, C corresponding to the same confidence coefficient α. This could be described by saying that the intervals δ_0 (E) are not *unnecessarily* broad.

The problem of determining the shortest system of confidence intervals is immediately reduced to that of finding appropriate regions of acceptance. In fact, using the Proposition I and II or the Corollary I expressed by (26), we may rewrite the condition (97) as follows :

$$P\{E \varepsilon A_0 (\theta'_1) | \theta_1^0\} \leq P\{E \varepsilon A (\theta'_1) | \theta_1^0\}, \quad \ldots\ldots \quad (98)$$

where $A_0 (\theta_1)$ and $A (\theta_1)$ denote the regions of acceptance leading to the systems of confidence intervals C_0 and C respectively.

If C_0 is the shortest system, then (98) should hold whatever θ_1^0 and θ'_1 and whatever the regions of acceptance $A (\theta_1)$, provided they correspond to the fixed confidence coefficient α. The condition (98) concerns the region of acceptance $A_0 (\theta'_1)$, and it must be combined with that expressed by the Proposition III, namely that

$$P\{E \varepsilon A_0 (\theta'_1) | \theta'_1\} = P\{E \varepsilon A (\theta'_1) | \theta'_1\} = \alpha, \quad \ldots\ldots \quad (99)$$

which must also hold for any θ'_1 and any values of the other parameters $\theta_2, \ldots \theta_l$.

We see that the problem of the shortest systems of confidence intervals corresponding to a confidence coefficient α is reduced to the following :

(1) Fix any value of $\theta_1 = \theta'_1$ and determine on the hyperplane $G (\theta'_1)$ a region $A (\theta'_1)$ similar to the sample space with regard to $\theta_2, \ldots \theta_l$ and of the size α.

(2) Out of all such regions $A (\theta'_1)$ choose the one, $A_0 (\theta'_1)$, for which the probability $P\{E \varepsilon A (\theta'_1) | \theta_1^0\}$, where θ_1^0 is any value of θ_1 different from θ'_1, is minimum.

(3) If the region $A_0 (\theta'_1)$ so found does not lose its property of minimizing $P\{E \varepsilon A (\theta'_1) | \theta_1^0\}$ when the value θ_1^0 is changed, and if the whole system of the regions $A_0 (\theta'_1)$ corresponding to all possible values of θ_1 satisfies the conditions (i)–(iv) of p. 354, then it may be used as the system of regions of acceptance and will

determine the shortest system of confidence intervals. The problem as described in (1) and (2) has already been considered in connexion with the theory of testing statistical hypotheses (NEYMAN and PEARSON, 1933) and its solution is known. However, it is also known that the region, $A_0(\theta'_1)$, satisfying the conditions (1) and (2) for a particular θ_1^0 does not always do so when that value of θ_1^0 is changed. It follows that the shortest systems of confidence intervals do not always exist. Still, they do exist occasionally. The reader acquainted with the joint paper mentioned will have no difficulty in checking that the confidence intervals determined by (61) and (62) in the case of the above Example I form the shortest system of confidence intervals. Applying the theory of the same paper, it is also easy to see that the confidence intervals defined by (90) and (91) with $b = 1 - \alpha$ form a system which is shortest of all those which could be constructed, using regions of acceptance belonging to the family based on the specific sufficient statistic x_1.

These, however, are rather rare cases. In order to emphasize this rareness, we shall prove the following proposition.

Proposition X

(1) If the probability law $p(E|\theta)$ of the X's, depending upon one parameter θ, is continuous in the whole sample space W and if at any point of this space it admits a continuous derivative with regard to θ not identically equal to zero, and admitting differentiation under the sign of the integral taken over W ;

(2) If $A(\theta')$ is a region in the sample space W and θ' and θ'' are two particular values of θ, such that

$$P\{E \varepsilon A(\theta')|\theta'\} = \alpha \ , \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (100)$$

and

$$P\{E \varepsilon A(\theta')|\theta''\} \leq P\{E \varepsilon A|\theta''\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (101)$$

where A is any other region in W such that $P\{E \varepsilon A|\theta'\} = \alpha$;

(3) If on the boundary of $A(\theta')$ there exists at least one point where $p(E|\theta')$ is not zero, then there must exist a third value of $\theta = \theta'''$, and a region B in W, such that

$$P\{E \varepsilon B|\theta'\} = \alpha \ . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (102)$$

$$P\{E \varepsilon A(\theta')|\theta'''\} > P\{E \varepsilon B|\theta'''\}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (103)$$

It will be noticed that the Proposition X means that if the probability law of the X's satisfies the condition (1), then the shortest system of confidence intervals generally do not exist. It follows also that in such cases the uniformly most powerful tests of hypotheses specifying the value of θ cannot exist.

We shall prove the Proposition X, starting with the assumption that it is not correct and that whatever the value θ''', either smaller or larger than θ', and whatever the region B satisfying (102) it follows that

$$P\{E \varepsilon A(\theta')|\theta'''\} \leq P\{E \varepsilon B|\theta'''\}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (104)$$

It is known (NEYMAN and PEARSON, 1933) that in such a case, whatever the sample point E′ within the region A (θ'), then for any θ,

$$p\,(E'|\theta) \leq k\,(\theta)\,p\,(E'|\theta'), \quad \ldots\ldots\ldots \quad (105)$$

where $k\,(\theta)$ depends only on θ and not on the x's. At any point, E″, outside A (θ') we should have

$$p\,(E''|(\theta) \geq k\,(\theta)\,p\,(E''|\theta'). \quad \ldots\ldots\ldots \quad (106)$$

Owing to the continuity of the probability law $p\,(E|\theta)$ we shall have at any point E‴ on the boundary of A (θ'_1)

$$p\,(E'''|\theta) = k\,(\theta)\,p\,(E'''|\theta'). \quad \ldots\ldots\ldots \quad (107)$$

We shall assume that $p\,(E'''|\theta') > 0$. As $p\,(E'''|\theta)$ admits a derivative with regard to θ, it follows that $k\,(\theta)$ must admit one. It follows also from (107) that if $\theta \longrightarrow \theta'$ then $k\,(\theta) \longrightarrow 1$. Differentiating (107) with regard to θ, and putting $\theta - \theta' = \Delta\theta$, we can write the following expansion of $k\,(\theta)$

$$\begin{aligned} k\,(\theta) &= 1 + \Delta\theta\, k'\,(\theta' + q\,\Delta\theta) \\ &= 1 + \Delta\theta\, p'\,(E'''|\theta' + q\,\Delta\theta)\, p^{-1}\,(E'''|\theta')\,, 0 < q < 1, \end{aligned} \quad (108)$$

where the dashes indicate differentiation with regard to θ. On the other hand, we can write also

$$p\,(E'|\theta) = p\,(E'|\theta'_1) + \Delta\theta\, p'\,(E'|\theta' + r\Delta\theta),\ 0 < r < 1. \ . \ . \quad (109)$$

Substituting (108) and (109) in (105) and rearranging, we get

$$\Delta\theta \left((p'\,(E'|\theta' + r\Delta\theta) - \frac{p'\,(E'''|\theta' + q\Delta\theta)\,p\,(E'|\theta')}{p\,(E'''|\theta')} \right) \leq 0, \quad . \quad (110)$$

and this inequality must hold good at any point E′ within A (θ') and for any value of $\Delta\theta$. It follows that

$$p'\,(E'|\theta') - \frac{p'\,(E'''|\theta')\,p\,(E'|\theta')}{p\,(E'''|\theta')} = 0 \quad \ldots\ldots \quad (111)$$

at any point E′ within A (θ'). In fact, if the expression in the left-hand side of (111) were not zero, then, owing to the continuity of $p'\,(E|\theta)$, for sufficiently small values of $\Delta\theta$, the expression in brackets in (110) would not be zero and would have a constant sign. As $\Delta\theta$ may be both positive and negative, the inequality (110) would not be satisfied. Using the inequality (106) holding good at any point outside A (θ') and repeating the above argument, we could easily find that (111) must hold good also outside A (θ') and therefore in the whole sample space W. Now it is easy to see that $p'\,(E|\theta')$ must be identically equal to zero, which contradicts the hypothesis (1) of the proposition X.

To show this we consider the integral

$$\int \dots \int_W p\,(\mathrm{E}|\theta)\,dx_1 \dots dx_n = 1. \quad \dots\dots\dots \quad (112)$$

Differentiating it with regard to θ and putting $\theta = \theta'$, we get

$$\int \dots \int_W p'\,(\mathrm{E}|\theta')\,dx_1 \dots dx_n = 0. \quad \dots\dots\dots \quad (113)$$

We can calculate $p'\,(\mathrm{E}|\theta')$ from (111) and substitute into (113). Using again (112) we find

$$\frac{p'\,(\mathrm{E}'''|\theta')}{p\,(\mathrm{E}'''|\theta')} = 0. \quad \dots\dots\dots\dots \quad (114)$$

Substituting this again in (111) we find $p'\,(\mathrm{E}|\theta'_1) = 0$, whatever the point E in W. This proves the Proposition X.

As the majority of probability laws with which we deal in practice, *e.g.*, the normal law, satisfy the conditions of Proposition X, it is seen that, for practical purposes, some other type of systems of confidence intervals is required, as the shortest systems generally do not exist.

(*b*) *One-sided Estimation*

The proof of the above proposition is based upon the circumstance that the left-hand side of the inequality (110) must not change its sign, while $\Delta\theta$ is both positive and negative.

It is therefore obvious that if it were for some reasons required to determine regions of acceptance $\mathrm{A}_0\,(\theta)$ satisfying the conditions

$$\mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}_0\,(\theta_1)|\theta_1\} = \alpha, \quad \dots\dots\dots \quad (112)$$

whatever the value of θ_1 and whatever the values of other unknown parameters involved in the probability law of the X's, and also the condition

$$\mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}_0\,(\theta'_1)|\theta''_1\} \leqq \mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}\,(\theta'_1)|\theta''_1\}, \quad \dots\dots \quad (113)$$

whatever any other region $\mathrm{A}\,(\theta'_1)$ satisfying (112) and whatever θ'_1 and θ''_1, *provided, however, the difference between them* $\theta'_1 - \theta''_1$, *is either always positive or always negative*, then the solution of this problem would exist more frequently than that of the problem of the shortest systems of confidence intervals.

The application of the regions of acceptance having the above properties is found useful in problems which may be called those of one-sided estimation. In frequent practical cases we are interested only in one limit which the value of the estimated parameter cannot exceed in one or in the other direction. When analysing seeds,

we ask for the minimum per cent. of germinating grains which it is possible to guarantee. When testing a new variety of cereals we are again interested in the minimum of gain in yield over the established standard which it is likely to give. In sampling manufactured products, the consumer will be interested to know the upper limit of the percentage defective which a given batch contains. Finally, in certain actuarial problems, we may be interested in the upper limit of mortality rate of a certain society group for which only a limited body of data is available.

In all these cases we are interested in the value of one parameter, say, θ_1, and it is desired to determine only one estimate of the same, either $\underline{\underline{\theta}}\,(E)$ or $\overline{\overline{\theta}}\,(E)$, which we shall call the unique lower and the unique upper estimate respectively. If θ_1 is the percentage of germinating seeds, we are interested in its lower estimate $\underline{\underline{\theta}}\,(E)$ so as to be able to state that $\underline{\underline{\theta}}\,(E) \leq \theta_1$, while the estimation of the upper bound $\overline{\overline{\theta}}\,(E)$ is of much less importance. On the other hand, if it is the question of the upper limit of mortality rate, θ_2, then we desire to make statements as to its value in the form $\theta_2 \leq \overline{\overline{\theta}}\,(E)$, etc.

These are the problems of one-sided estimation, and it is easy to see that their most satisfactory solution depends upon the possibility of constructing regions of acceptance satisfying (1) and (2), the latter with the restriction that the sign of the difference $\theta'_1 - \theta''_1$ is constant.

The two problems of the unique lower and the unique upper estimates are very similar, so that it will be sufficient to treat only one of them, *e.g.*, the first. Suppose, then, that we are interested in the unique lower estimate $\underline{\underline{\theta}}\,(E)$ of a parameter θ_1. Treating the problem from the point of view of confidence intervals, we desire to define a function $\underline{\underline{\theta}}\,(E)$ of the sample point E such that whatever may be the true value θ_1^0, of θ_1, the probability

$$P\{\underline{\underline{\theta}}\,(E) \leq \theta_1^0 | \theta_1^0\} = \alpha \quad \ldots\ldots\ldots \quad (114)$$

where α is the chosen confidence coefficient. Repeating the reasonings of the preceding sections, we find that this problem is equivalent with that of choosing appropriate regions of acceptance and that there is an infinity of solutions. Let us now specify the properties of a solution which would make it more desirable than any other.

For that purpose denote by θ_1^0 the unknown true value of θ_1 and by θ'_1 and θ''_1 any two other values such that

$$\theta'_1 < \theta_1^0 < \theta''_1. \quad \ldots\ldots\ldots\ldots \quad (115)$$

It is obvious that if we are interested only in the unique lower estimate of θ_1 and want the probability of $\underline{\underline{\theta}}\,(E)$ falling short of the true value θ_1^0 to be equal to α, we should not mind $\underline{\underline{\theta}}\,(E)$ being smaller than θ''_1. Therefore, when choosing the function $\underline{\underline{\theta}}\,(E)$, we should not formulate any restriction concerning its satisfying the inequality $\underline{\underline{\theta}}\,(E) < \theta''_1$, provided the equation (114) is satisfied. The position with regard to θ'_1 is different. If $\underline{\underline{\theta}}\,(E)$ happens to be smaller than θ'_1, then it will also be

smaller than θ_1^0 and our statement concerning the value of θ_1 based on $\underline{\underline{\theta}}(E)$ will be correct. However, it would also be correct if, say,

$$\underline{\underline{\theta}}(E) = \tfrac{1}{2}(\theta'_1 + \theta_1^0) > \theta'_1, \quad \ldots\ldots\ldots \quad (116)$$

and in such a case it would be more accurate and would undoubtedly be judged more desirable. Generalizing the above conclusion, we could say that whenever we are interested in the unique lower estimate $\underline{\underline{\theta}}(E)$ of a parameter θ_1, we should require it to have the property that whatever $\theta'_1 < \theta_1^0$, the chance of $\underline{\underline{\theta}}(E)$ falling short of θ'_1 should be as small as possible, thus

$$P\{\underline{\underline{\theta}}(E) < \theta'_1 | \theta_1^0\} = \text{minimum} \quad \ldots\ldots\ldots \quad (117)$$

for all values of θ'_1 and θ_1^0 such that $\theta'_1 < \theta_1^0$. This condition implies that the region of acceptance $A_0(\theta'_1)$ corresponding to any value of $\theta_1 = \theta'_1$ should have the property

$$P\{E \varepsilon A_0(\theta'_1) | \theta_1^0\} \leqq P\{E \varepsilon A | \theta_1^0\}, \quad \ldots\ldots\ldots \quad (118)$$

whatever $\theta_1^0 > \theta'_1$ and whatever any other region A such that

$$P\{E \varepsilon A | \theta'_1\} = P\{E \varepsilon A_0(\theta'_1) | \theta'_1\} = \alpha. \quad \ldots\ldots\ldots \quad (119)$$

Similarly, if it were desired to find the unique upper estimate $\overline{\overline{\theta}}(E)$ of θ_1, the most desirable solution would be determined by the regions of acceptance, $A^0(\theta_1)$ such that

$$P\{E \varepsilon A^0(\theta'_1) | \theta_1^0\} \leqq P\{E \varepsilon A | \theta_1^0\}, \quad \ldots\ldots\ldots \quad (120)$$

whatever $\theta_1^0 < \theta'_1$ and whatever the region A satisfying (119).

If unique estimates determined by (118) and (119) or (120) and (119) exist, they will be called the best one-sided estimates of θ_1.

Following the recent results (NEYMAN and PEARSON, 1933, 1936, *a*) concerning the theory of testing hypotheses, it is easy to establish formulae giving the best one-sided estimates in many important problems. Of these I shall mention one.

(*c*) *Example III*

Consider the case where the probability law of the X's is normal

$$p(E|\xi\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\Sigma(x_i-\xi)^2}{2\sigma^2}} \quad \ldots\ldots\ldots \quad (121)$$

with unknown ξ and σ and where it is desired to estimate ξ. Following the lines indicated, it is easily found that the best one-sided estimates of ξ are given by

$$\left.\begin{aligned} \overline{\overline{\xi}}(E) &= \bar{x} + ts \\ \underline{\underline{\xi}}(E) &= \bar{x} - ts \end{aligned}\right\}, \quad \ldots\ldots\ldots\ldots \quad (122)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)} \quad \ldots\ldots\ldots \quad (123)$$

and t may be taken from Fisher's tables corresponding to $P = 2(1 - \alpha)$.*

(*d*) *Short Unbiassed Systems of Confidence Intervals*

We must now consider the important case where we are interested in the two-sided estimation in which the probability law of the X's is highly regular so that, owing to Proposition X, the shortest systems of confidence intervals do not exist. We must formulate the properties of confidence intervals which could be considered as particularly satisfactory.

We shall start with the obvious remark that, if possible, the value of the estimated parameter which in a particular case happens to be true, should be covered by the confidence interval more frequently than any other value.

Alternatively, we may express this by saying that for any given value of $\theta_1 = \theta_1^0$ the probability of its being covered by the confidence interval δ (E) should be greatest when θ_1^0 happens to be the true value of θ_1. Therefore, whatever $\theta'_1 \neq \theta_1^0$, it must be

$$P\{\delta(E)\, C\theta_1^0 | \theta_1^0\} = \alpha \geq P\{\delta(E)\, C\theta_1^0 | \theta'_1\}. \quad \ldots\ldots \quad (124)$$

We may express this in still another manner, considering the probability of θ_1^0 being covered by the confidence interval δ (E) as a function of that value of θ_1 which happens to be true,

$$P\{\delta(E)\, C\theta_1^0 | \theta_1\} = f(\theta_1). \quad \ldots\ldots\ldots\ldots \quad (125)$$

The formula (124) requires that the function (125) should be maximum for $\theta_1 = \theta_1^0$ and that that maximum should be equal to α.

It seems to be obvious that if there are many systems of confidence intervals in which, whatever θ_1^0, the probability (125) considered as a function of θ_1, is maximum for $\theta_1 = \theta_1^0$, we should choose the system by which this maximum is the steepest, so that, while the true value of θ_1 is being shifted away from θ_1^0, the chance of θ_1^0 being covered by δ (E) diminishes in the quickest way.

These conditions may now be expressed in terms of equivalent conditions concerning the regions of acceptance.

* The properties of the formulae (122) giving the best one-sided estimates of ξ were found by the author in about 1930. Subsequently, these properties, together with an outline of the theory of estimation, were included in his lectures first given at the University of Warsaw, then, from 1934, at the University College, London, and also in a course of lectures at the University of Paris in January, 1936. References to these formulae may be found both in Polish and English statistical literature. *See* for instance: (1) W. PYTKOWSKI: "The Dependence of the Income of Small Farms upon their Area, the Outlay and the Capital Invested in Cows". Warsaw, 1932. *See* particularly pp. 28–29; (2) CLOPPER and PEARSON (1934).

Let $A(\theta_1^0)$ be a region of acceptance corresponding to some value θ_1^0 of θ_1, so that

$$P\{E \varepsilon A(\theta_1^0) | \theta_1^0\} = \alpha \quad \ldots \quad (126)$$

whatever θ_1^0. We have

$$f(\theta_1) = P\{\delta(E) C\theta_1^0 | \theta_1\} = P\{E \varepsilon A(\theta_1^0) | \theta_1\} \quad \ldots \quad (127)$$

and the above conditions concerning the confidence intervals appear to be equivalent with the condition that the right-hand side of (127), considered as a function of θ_1, should be a maximum for $\theta_1 = \theta_1^0$ and that this maximum should be as sharp as possible.

In cases where the elementary probability law of the X's, integrated over any region, admits two differentiations with regard to θ_1 under the integral sign, this leads to the following :

Whatever θ_1^0, and *whatever the values of other unknown parameters*, $\theta_2, \theta_3, \ldots \theta_l$,

$$\left.\frac{\partial P\{E \varepsilon A(\theta_1^0) | \theta_1\}}{\partial \theta_1}\right|_{\theta_1 = \theta_1^0} \equiv \int \ldots \int_{A(\theta_1^0)} p'(E | \theta_1^0, \ldots \theta_l)\, dx_1 \ldots dx_n \equiv 0 \quad \ldots \quad (128)$$

$$\left.\frac{\partial^2 P\{E \varepsilon A(\theta_1^0) | \theta_1\}}{\partial \theta_1^2}\right|_{\theta_1 = \theta_1^0} \equiv \int \ldots \int_{A(\theta_1^0)} p''(E | \theta_1^0, \ldots \theta_l)\, dx_1 \ldots dx_n = \text{minimum}, \quad (129)$$

where p' and p'' denote the derivatives with regard to θ_1.

The system of confidence intervals having the above properties will be called the short unbiassed system. The possibility of determining such systems depends on the possibility of determining the regions of acceptance satisfying (126), (128), and (129). This problem has been recently dealt with in the case where the number of the unknown parameters involved in the probability law of the X's is equal to one (NEYMAN and PEARSON, 1936, *a*) and to two (NEYMAN, 1935, *b*).

In such cases as treated in the papers referred to, the construction of the short unbiassed systems of confidence intervals does not present any difficulties.

In particular, if the probability law of the X's is as in (121), then the short unbiassed system of the confidence intervals for ξ is given by the formula

$$\bar{x} - ts \leq \xi \leq \bar{x} + ts \quad \ldots \quad (130)$$

where t should be taken from Fisher's tables for $P = 1 - \alpha$.

IV—SUMMARY

The main problem treated in this paper is that of confidence limits and of confidence intervals and may be briefly described as follows. Let $p(x_1, \ldots x_n | \theta_1, \theta_2, \ldots \theta_l) = p(E | \theta_1, \ldots \theta_l)$ be the elementary probability law of n random variables $x_1 \ldots x_n$ depending on l constant parameters $\theta_1, \theta_2, \ldots \theta_l$. The letter E stands here for $x_1, \ldots x_n$. Suppose that the analytical nature of $p(x_1, \ldots x_n | \theta_1, \ldots \theta_l)$ is known but the values of the parameters $\theta_1, \ldots \theta_l$ are unknown. It is required to determine two

single-valued functions of the x's, $\underline{\theta}$ (E) and $\bar{\theta}$ (E) having the property that, whatever the values of the θ's, say $\theta'_1, \theta'_2, \ldots \theta'_l$, the probability of $\underline{\theta}$ (E) falling short of θ'_1 and at the same time of $\bar{\theta}$ (E) exceeding θ'_1 is equal to a number α fixed in advance so that $0 < \alpha < 1$,

$$P\{\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\} = \alpha. \quad . \quad . \quad . \quad . \quad . \quad (131)$$

It is essential to notice that in this problem the probability refers to the values of $\underline{\theta}$ (E) and $\bar{\theta}$(E) which, being single-valued functions of the x's, are random variables. θ'_1 being a constant, the left-hand side of (131) *does not* represent the probability of θ'_1 falling within some fixed limits.

The functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are called the confidence limits for θ_1 and the interval ($\underline{\theta}$ (E) , $\bar{\theta}$ (E)) the confidence interval corresponding to the confidence coefficient α.

The problem thus stated has been completely solved for the case where $l = 1$, and it is found to possess an infinity of solutions. If $l \geq 2$ the solution obtained is limited to the case where there exists a sufficient set of statistics for $\theta_2, \theta_3, \ldots \theta_l$ and then again there is an infinity of solutions.

Methods were indicated by which it is possible to find among all possible solutions of the problem the one giving the confidence intervals which are shorter (in a sense defined in the text) than those corresponding to any other solution.

The confidence limits $\underline{\theta}$ (E) and $\bar{\theta}$ (E) may be looked upon as giving a solution of the statistical problem of estimating θ_1 independent of any knowledge of probabilities *a priori*. Once $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined corresponding to a value of α close to unity, say $\alpha = 0{\cdot}99$, the statistician desiring to estimate θ_1 may be recommended (1) to observe the values of the random variables $x_1, \ldots x_n$, (2) to calculate the corresponding values of $\underline{\theta}$ (E) and $\bar{\theta}$ (E), and (3) to state that the value of the parameter θ_1 is within the limits $\underline{\theta}\,(E) \leq \theta_1 \leq \bar{\theta}\,(E)$.

The justification of this recommendation lies in the fact that the three steps described are equivalent to a random experiment which may result either in a correct or in an erroneous statement concerning the value of θ_1, the probability of a correct statement being equal to $\alpha = 0{\cdot}99$. Thus the statistician following the above recommendation is in a position comparable with that of a game of chance with the probability of winning being equal to $\alpha = 0{\cdot}99$.

The method followed to determine the confidence limits for a single parameter permits an obvious generalization for the case where the number of parameters to be estimated simultaneously is greater than one.

Three previous publications concerning the confidence intervals for which I am either partly or wholly responsible (NEYMAN, 1934, MATUSZEWSKI, NEYMAN, and SUPIŃSKA, 1935, NEYMAN, 1935, *c*) refer to the simplest case where the number of random variables and that of the parameters to be estimated are equal to unity. The problem considered here is therefore far more general and also it is treated differently. Previously, the parameters to be estimated were considered as random variables following an arbitrary probability law which could be continuous or not and, even,

could reduce to unity just for one particular value of the parameter, being zero elsewhere. This arbitrariness of the probability law of the parameters served as an excuse, but the very assumption of its existence seemed to be an artificiality from which the present method of approach is entirely free.

Subsidiary results obtained include a method of constructing similar regions which is more general than the one known previously and the Proposition X bearing on the theory of testing hypotheses. It emphasizes the rareness of cases where there exists a uniformly most powerful test.

V—References

Borel, É. 1910 " Eléments de la Theorie des Probabilités," Paris.

—— 1925 " Principes et formules classiques," 1 fasc. du tome I du " Traité du Calcule des Probabilités et des ses Applications," Paris.

—— 1926 " Applications à l'Arithmétique," Ibidem fasc. 1, tome II.

Bortkiewicz, L. v. 1917 " Die Iterationen." Berlin.

Clopper, C. J., and Pearson, E. S. 1934 " Biometrika," **26**, 404–413.

Darmois, G. 1936 " Méthodes d'estimation," Paris.

Doob, J. L. 1934 ' Trans. Amer. Math. Soc.,' **36**, 759.

Dugué, D. 1936 ' C.R. Acad. Sci. Paris,' **193**, 452, 1732.

Fisher, R. A. 1925 ' Proc. Camb. Phil. Soc.,' **22**, 700–725.

Fréchet, M. 1937 " Recherches théoriques modernes sur le calcul des probabilités. Traité du Calcul des Probabilités et de ses Applications," Fasc. 3, tome 1, Paris.

Hopf, E. 1934 ' J. Math, Phys. Mass,' **13**.

Hotelling, H. 1932 ' Trans. Amer. Math. Soc.,' **32**, 847–859.

Jeffreys, H. 1931 " Scientific Inference," 1935.

Kolmogoroff, A. 1933 " Grundbegriffe der Wahrscheinlichkeitsrechnung," Berlin.

Lévy, P. 1925 " Calcul des Probabilités," Paris.

Łomnicki, Z., and Ulam, S. 1934 ' Fund. Math.,' **23**, 237–238.

Markoff, A. A. 1923 " Calculus of Probability " (Russian ed. iv), Moscow.

Matuszewski, T., Neyman, J., and Supińska, J. 1935 ' Supplement to J. Roy. Stat. Soc.,' **1**, 63–82.

Neyman, J., and Pearson, E. S. 1933 ' Phil. Trans.,' A, **231**.

—— 1936, *a* ' Stat. Res. Mem.,' **1**, 1–37.

—— 1936, *b* ' Stat. Res. Mem.,' **1**, 113–137.

Neyman, J. 1935, *a* ' G. Inst. Ital. Attuari,' **6**, 320.

—— 1935, *b* ' Bull. Soc. Math. Fr.,' **63**, 248–266.

Neyman, J. 1935, *c* ' Ann. Math. Stat.,' **6**, 111–116.

—— 1934 ' J. Roy. Stat. Soc.,' **97**, 589–593.

Pearson, K. 1895 ' Phil. Trans.,' **187**, 253–318.

Tippett, L. H. C. 1927 " Tracts for Computers," No. 15, Camb. Univ. Press.