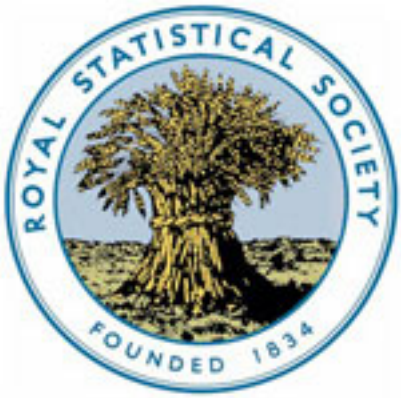


On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection.

By Jerzy Neyman

(Biometric Laboratory, Nencki Institute, Soc. Sci. Lit. Varsoviensis, Warsaw).

[Read before the Royal Statistical Society, June 19th, 1934, the President, the Rt. Hon. Lord Meston of Agra and Dunottar, K.C.S.I., LL.D., in the Chair.]

CONTENTS.

I. Introductory.

Owing to the work of the International Statistical Institute,* and perhaps still more to personal achievements of Professor A. L. Bowley, the theory and the possibility of practical applications of the representative method has attracted the attention of many statisticians in different countries. Very probably this popularity of the representative method is also partly due to the general crisis, to the scarcity of money and to the necessity of carrying out statistical investigations connected with social life in a somewhat hasty way. The results are wanted in some few months, sometimes in a few weeks after the beginning of the work, and there is neither time nor money for an exhaustive research.

But I think that if practical statistics has acquired something

* See "The Report on the Representative Method in Statistics" by A. Jensen, *Bull. Inst. Intern. Stat.*, XXII. 1ère Livr.

valuable in the representative method, this is due primarily to Professor A. L. Bowley, who not only was one of the first to apply this method in practice,* but also wrote a very fundamental memoir † giving the theory of the method. Since then the representative method has been often applied in different countries and for different purposes.

My chief topic being the theory of the representative method, I shall not go into its history and shall not quote the examples of its practical application however important—unless I find that their consideration might be useful as an illustration of some points of the theory.

There are two different aspects of the representative method. One of them is called the method of random sampling and the other the method of purposive selection. This is a division into two very broad groups and each of these may be further subdivided. The two kinds of method were discussed by A. L. Bowley in his book, in which they are treated as it were on equal terms, as being equally to be recommended. Much the same attitude has been expressed in the Report of the Commission appointed by the International Statistical Institute for the purpose of studying the application of the Representative Method in Statistics.‡ The Report says: "In the selection of that part of the material which is to be the object of direct investigation, one or the other of the following two principles can be adopted: in certain instances it will be possible to make use of a combination of both principles. The one principle is characterized by the fact that the units which are to be included in the sample are selected at random. This method is only applicable where the circumstances make it possible to give every single unit an equal chance of inclusion in the sample. The other principle consists in the samples being made up by purposive selection of groups of units which it is presumed will give the sample the same characteristics as the whole. There will be especial reason for preferring this method, where the material differs with respect to composition from the kind of material which is the basis of the experience of games of chance, and where it is therefore difficult or even impossible to comply with the aforesaid condition for the application of selection at random. Each of these two methods has certain advantages and certain defects. . . ."

This was published in 1926. In November of the same year

* A. L. Bowley: "Working Class Households in Reading." *J.R.S.S.*, June, 1913.

† A. L. Bowley: "Measurement of the Precision Attained in Sampling." Memorandum published by the Int. Stat. Inst., *Bull. Int. Stat. Inst.*, Vol. XXII. 1ère Livr.

‡ *Bull. Int. Stat. Inst.*, XXII. 1ère Livr. p. 376.

the Italian statisticians C. Gini and L. Galvani were faced with the problem of the choice between the two principles of sampling, when they undertook to select a sample from the data of the Italian General Census of 1921. All the data were already worked out and published and the original sheets containing information about individual families were to be destroyed. In order to make possible any further research, the need for which might be felt in the future, it was decided to keep for a longer time a fairly large sample of the census data, amounting to about 15 per cent. of the same.

The chief purpose of the work is stated by the authors as follows : * " To obtain a sample which would be representative of the whole country with respect to its chief demographic, social, economic and geographic characteristics."

At the beginning of the work the original data were already sorted by provinces, districts (circondarî) and communes, and the authors state that the easiest method of obtaining the sample was to select data in accordance with the division of the country in administrative units. As the purpose of the sample was among others to allow local comparisons to be made in the future, the authors expressed the view that the selection of the sample, taking administrative units as elements, was the only possible one.

For various reasons, which, however, the authors do not describe, it was impossible to take as an element of sampling an administrative unit smaller than a commune. They did not, however, think it satisfactory to use communes as units of selection because (p. 3 *loc. cit.*) their large number (8,354) would make it difficult to apply the method of purposive selection. So finally the authors fixed districts (circondarî) to serve as units of sampling. The total number of the districts in which Italy is divided amounts to 214. The number of the districts to be included in the sample was 29, that is to say, about 13·5 per cent. of the total number of districts.

Having thus fixed the units of selection, the authors proceed to the choice of the principle of sampling : should it be random sampling or purposive selection? To solve this dilemma they calculate the probability, π, that the mean income of persons included in a random sample of $k = 29$ districts drawn from their universe of $K = 214$ districts will differ from its universe-value by not more than 1·5 per cent. The approximate value of this probability being very small, about $\pi = \cdot 08$, the authors decided that the principle of sampling to choose was that of purposive selection.†

The quotation from the Report of the Commission of the Inter-

* *Annali di Statistica*, Ser. VI. Vol. IV. p. 1. 1929.

† It may be noted, however, that the choice of the principle seems to have been predetermined by the previous choice of the unit of sampling.

national Statistical Institute and the choice of the principle of sampling adopted by the Italian statisticians, suggest that the idea of a certain equivalency of both principles of random sampling and purposive selection is a rather common one. As the theory of purposive selection seems to have been extensively presented only in the two papers mentioned, while that of random sampling has been discussed probably by more than a hundred authors, it seems justifiable to consider carefully the basic assumptions underlying the former. This is what I intend to do in the present paper. The theoretical considerations will then be illustrated on practical results obtained by Gini and Galvani, and also on results of another recent investigation, carried out in Warsaw, in which the representative method was used. As a result of this discussion it may be that the general confidence which has been placed in the method of purposive selection will be somewhat diminished.

II. Mathematical Theories underlying the Representative Method.

1. *The Theory of Probabilities* a posteriori *and the work of R. A. Fisher.*

Obviously the problem of the representative method is *par excellence* the problem of statistical estimation. We are interested in characteristics of a certain population, say π, which it is either impossible or at least very difficult to study in detail, and we try to estimate these characteristics basing our judgment on the sample. Until recently it has been usually assumed that the accurate solution of such a problem requires the knowledge of probabilities *a priori* attached to different admissible hypotheses concerning the values of the collective characters * of the population π. Accordingly, the memoir of A. L. Bowley may be regarded as divided into two parts. Each question is treated from two points of view : (*a*) The population π is supposed to be known; the question to be answered is : what could be the samples from this population ? (*b*) We know the sample and are concerned with the probabilities *a posteriori* to be ascribed to different hypotheses concerning the population.

In sections which I classify as (*a*) we are on the safe ground of classical theory of probability, reducible to the theory of combinations.†

In sections (*b*), however, we are met with conclusions based,

* This is a translation of the terminology used by Bruns and Orzecki. Any characteristics of the population or sample is a collective character.

† In this respect I should like to call attention to the remarkable paper of the late L. March published in *Metron*, Vol. VI. There is practically no question of probabilities and many classical theorems of this theory are reduced to the theory of combinations.

inter alia, on some quite arbitrary hypotheses concerning the probabilities *a priori*, and Professor Bowley accompanies his results with the following remark: "It is to be emphasized that the inference thus formulated is based on assumptions that are difficult to verify and which are not applicable in all cases."

However, since Bowley's book was written, an approach to problems of this type has been suggested by Professor R. A. Fisher which removes the difficulties involved in the lack of knowledge of the *a priori* probability law.* Unfortunately the papers referred to have been misunderstood and the validity of statements they contain formally questioned. This I think is due largely to the very condensed form of explaining ideas used by R. A. Fisher, and perhaps also to a somewhat difficult method of attacking the problem. Avoiding the necessity of appeals to the somewhat vague statements based on probabilities *a posteriori*, Fisher's theory becomes, I think, the very basis of the theory of representative method. In Note I in the Appendix I have described its main lines in a way somewhat different from that followed by Fisher.

The possibility of solving the problems of statistical estimation independently from any knowledge of the *a priori* probability laws, discovered by R. A. Fisher, makes it superfluous to make any appeals to the Bayes' theorem.

The whole procedure consists really in solving the problems which Professor Bowley termed direct problems: given a hypothetical population, to find the distribution of certain characters in repeated samples. If this problem is solved, then the solution of the other problem, which takes the place of the problem of inverse probability, can be shown to follow.

The form of this solution consists in determining certain intervals, which I propose to call the confidence intervals (see Note I), in which we may assume are contained the values of the estimated characters of the population, the probability of an error in a statement of this sort being equal to or less than $1 - \varepsilon$, where ε is any number $0 < \varepsilon < 1$, chosen in advance. The number ε I call the confidence coefficient. It is important to note that the methods of estimating, particularly in the case of large samples, resulting from the work of Fisher, are often precisely the same as those which are already in common use. Thus the new solution of the problems of estimation consists mainly in a rigorous justification of what has been generally considered correct more or less on intuitive grounds.†

* R. A. Fisher: *Proc. Camb. Phil. Soc.*, Vol. XXVI, Part 4, Vol. XXVIII, Part 3, and *Proc. Roy. Soc.*, A. Vol. CXXXIX.

† I regret that the necessarily limited size of the paper does not allow me to go into the details of this important question. It has been largely studied by R. A. Fisher. His results in this respect form a theory which he calls the

Here I should like to quote the words of Laplace, that the theory of probability is in fact but the good common sense which is reduced to formulæ. It is able to express in exact terms what the sound minds feel by a sort of instinct, sometimes without being able to give good reasons for their beliefs.

2. *The Choice of the Estimates.*

However, it may be observed that there remains the question of the choice of the collective characters of the samples which would be most suitable for the purpose of the construction of confidence intervals and thus for the purposes of estimation. The requirements with regard to these characters in practical statistics could be formulated as follows:

1. They must follow a frequency distribution which is already tabled or may be easily calculated.

2. The resulting confidence intervals should be as narrow as possible.

The first of these requirements is somewhat opportunistic, but I believe as far as the practical work is concerned this condition should be borne in mind.*

Collective characters of the samples which satisfy both conditions quoted above and which may be used in the most common cases, are supplied by the elegant method of A. A. Markoff,† used by him when

---

Theory of Estimation. The above-mentioned problems of confidence intervals are considered by R. A. Fisher as something like an additional chapter to the Theory of Estimation, being perhaps of minor importance. However, I do not agree in this respect with Professor Fisher. I am inclined to think that the importance of his achievements in the two fields is in a relation which is inverse to what he thinks himself. The solution of the problem which I described as the problem of confidence intervals has been sought by the greatest minds since the work of Bayes 150 years ago. Any recent book on the theory of probability includes large sections concerning this problem. These sections are crowded with all sorts of "paradoxes," etc. The present solution means, I think, not less than a revolution in *the theory* of statistics. On the other hand, the problem of the choice of estimates has—as far as I can see—mainly a *practical* importance. If this is not properly solved (granting that the problem of confidence intervals has been solved correctly) the resulting confidence intervals will be unnecessarily broad, but our statements about the values of estimated collective characters will still remain correct. Thus I think that the problems of the choice of the estimates are rather the technical problems, which, of course, are extremely important from the point of view of practical work, but the importance of which cannot be compared with the importance of the other results of R. A. Fisher, concerning the very basis of the modern statistical theory. These are, of course, "qualifying judgments," which may be defended and may be attacked, but which anyone may accept or reject, according to his personal point of view and the perspective on the theory of statistics.

* The position is a different one if we consider the question from the point of view of the theory. Here I have to mention the important papers of R. A. Fisher on the theory of likelihood.

† A. A. Markoff: Calculus of Probabilities. Russian. Edition IV, Moscow 1923. There was a German edition of this book, Leipzig 1912, actually out of print.

dealing with the theory of least squares. The method is not a new one, but as it was published in Russian it is not generally known.* This method, combined with some results of R. A. Fisher and of E. S. Pearson concerning the extension of " Student's " distribution allows us to build up the theory of different aspects of representative method to the last details.

Suppose θ is a certain collective character of a population π and

$$x_1, x_2, \ldots x_n \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

is the sample from this population. We shall say that a function of these x's, say

$$\theta' = \theta'(x_1, x_2, \ldots x_n) \quad . \quad . \quad . \quad . \quad . \quad (2)$$

is a " mathematical expectation estimate " † of θ, if the mean value of θ' in repeated samples is equal to θ. Further, we shall say that the estimate θ' is the best linear estimate of θ if it is linear with regard to x's, *i.e.*

$$\theta' = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n + \lambda_0 \quad . \quad . \quad (3)$$

and if its standard error is less than the standard error of any other linear estimate of θ.

Of course, in using the words " best estimate " I do not mean that the estimate defined has unequivocable advantages over all others. This is only a convention and, as long as the definition is borne in mind, will not cause any misunderstanding. Still, the best linear estimates have some important advantages :

1. If n be large, their distribution practically always follows closely the normal law of frequency. This is important, as in applying the representative method in social and economic statistics we are commonly dealing with very large samples.

2. In most cases they are easily found by applying Markoff's method.

3. The same method provides us with the estimate of their standard errors.

4. If the estimate θ' of θ is a linear estimate, and if μ is the estimate of its standard error, then, in cases when the sampled population is normally distributed, the ratio

$$t = \frac{\theta' - \theta}{\mu} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

follows the " Student's " distribution, which is dependent only upon the size of the sample. This is the result due to R. A.

* I doubt, for example, whether it was known to Bowley and to Gini and Galvani when they wrote their papers.

† Only the estimates of this kind will we consider below.

Fisher. Moreover, R. A. Fisher has provided tables giving the values of t such that the probability of their being exceeded by $|\theta' - \theta|/\mu$ has definite values such as ·01, ·02, . . . etc. This table * was published long before any paper dealing with the solution of the problem of estimation independent of the probabilities *a priori*. However, this solution is already contained in the table. In fact it leads directly to the construction of the confidence intervals. Suppose the confidence coefficient chosen is $\varepsilon = ·99$. Obtain from Fisher's table the value of t, say t_ϵ, corresponding to the size of the sample we deal with and to a probability of its being exceeded by $|\theta' - \theta|/\mu$ equal to $1 - \varepsilon = ·01$. It may then be easily shown that the confidence interval, corresponding to the coefficient $\varepsilon = ·99$ and to the observed values of θ' and μ, will be given by the inequality

$$\theta' - \mu t_\epsilon \leq \theta \leq \theta' + \mu t_\epsilon \quad . \quad . \quad . \quad . \quad . \quad (5)$$

5. The previous statement is rigorously true if the distribution of the x's is normal. But, as it has been experimentally shown by E. S. Pearson,† the above result is very approximately true for various linear estimates by fairly skew distributions, provided the sample dealt with is not exceedingly small, say not smaller than of 15 individuals. Obviously, when applying the representative method to social problems this is a limitation of no importance. In fact, if the samples are very large, the best linear estimates follow the normal law of frequency, and the multiplier t_ϵ in the formula giving the confidence interval may be found from any table of the normal integral.‡

The above properties of the linear estimates make them exceedingly valuable from the point of view of their use in applying the representative method. I proceed now to the Markoff method of finding the best linear estimates.

This may be applied under the following conditions, which are frequently satisfied in practical work.

Suppose we are dealing with k populations,

$$\pi_1, \pi_2, \ldots \pi_k \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

from which we may draw random samples. Let

$$x_{i1}, x_{i2}, \ldots x_{ini} \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

be a sample, Σ_i, of n_i individuals randomly drawn (with replacement or not) from the population π_i. Let A_i be the mean of the

* R. A. Fisher: *Statistical Methods for Research Workers*, London, 1932, Edition IV.

† This *Journal*, Vol. XCVI, Part I.

‡ For example, Table I of the Pearson's Tables for Statisticians and Biometricians, Part I, may be used.

population π_i. We have now to make some assumption about the variances, σ_i^2, of the populations π_i. The actual knowledge of these variances is not required. But we must know numbers which are proportional to σ_i^2. Thus we shall assume that

$$\sigma_i^2 = \frac{\sigma_0^2}{P_i} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

σ_0^2 being an unknown factor, and P_i a known number.* It would be a special case of the above conditions if it were known that

$$\sigma_1 = \sigma_2 = \ldots = \sigma_k \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

the common value of the σ's being unknown.

Suppose now we are interested in the values of one or several collective characters of the populations, π_i, each of them being a linear function of the means of these populations, say

$$\theta_j = a_{j1}A_1 + a_{j2}A_2 + \ldots + a_{jk}A_k \quad . \quad . \quad (10)$$

where the a's are some known coefficients. Markoff gives now the method of finding linear functions of the x's determined by samples from all the populations, namely,

$$\begin{aligned} \theta'_j = {} & \lambda_{11}x_{11} + \lambda_{12}x_{12} + \ldots + \lambda_{1n_1}x_{1n_1} + \\ & + \ldots\ldots\ldots\ldots\ldots + \\ & \ldots\ldots\ldots\ldots\ldots + \\ & + \lambda_{k1}x_{k1} + \lambda_{k2}x_{k2} + \ldots + \lambda_{kn_k}x_{kn_k} \end{aligned} \quad . \quad (11)$$

such, *that whatever the value of unknown* θ_j :

(a) Mean θ'_j in repeated samples $= \theta_j$.

(b) Standard error of θ'_j is less than that of any other linear function, satisfying (a).

The details concerning this method are given in Note II of the Appendix.

It is worth considering the statistical meaning of the two conditions (a), (b), when combined with the fact that if the number of observations is large, the distribution of θ' in repeated sampling tends to be, and for practical purposes *is* actually normal. The condition (a) means that the most frequent values of θ' will be those close to θ. Therefore, if ψ is some linear function of the x's, which does not satisfy the condition (a), but instead the condition,

Mean ψ in repeated samples $= \theta + \Delta$, (say),

then, using ψ as an estimate of θ, we should commit systematic errors, which most frequently would be near Δ. Such estimates as ψ are called biased.

The condition (b) assures us that when using θ''s as estimates of

* Sometimes, in special problems, even this knowledge is not required.

0's, we shall get confidence intervals corresponding to a definite confidence coefficient, narrower than those obtained using any other linear estimate. In other words, using linear estimates satisfying the conditions (*a*) and (*b*) we may be sure that we shall not commit systematic errors, and that the accuracy of the estimate will be the greatest.

III. Different Aspects of the Representative Method.

We may now proceed to consider the two aspects of the representative method.

1. *The Method of Random Sampling.*

The method of random sampling consists, as it is known, in taking at random elements from the population which it is intended to study. The elements compose a sample which is then studied. The results form the basis for conclusions concerning the population. The nature of the population is arbitrary. But we shall be concerned with populations of inhabitants of some country, town, etc. Let us denote this population by Π. Its elements will be single individuals, of which we shall consider a certain character x, which may be measurable or not (*i.e.* an attribute). Suppose we want to estimate the average value of the character x, say X, in all individuals forming the population Π. It is obvious that in the case where x is an attribute, which may be possessed or not by the individuals of the population, its numerical value in these individuals will be 0 or 1, and its mean value X will be the proportion of the individuals having actually the attribute x.

The method of random sampling may be of several types :

(*a*) The sample, Σ, which we draw to estimate X is obtained by taking at random single individuals from the population Π. The method of sampling may be either that with replacement or not. This type has been called by Professor Bowley that of unrestricted sampling.

(*b*) Before drawing the random sample from the population Π this is divided into several " strata," say

$$\Pi_1, \Pi_2, \ldots \Pi_k \quad . \quad . \quad . \quad . \quad . \quad (12)$$

and the sample Σ is composed of k partial samples, say

$$\Sigma_1, \Sigma_2, \ldots \Sigma_k \quad . \quad . \quad . \quad . \quad . \quad (13)$$

each being drawn (with replacement or not) from one or other of the strata. This method has been called by Professor Bowley the method of stratified sampling. Professor Bowley considered only the case when the sizes, say, m'_i, of the partial samples are pro-

portionate to the sizes of corresponding strata. I do not think that this restriction is necessary and shall consider the case when the sizes of the strata, say

$$M'_1, M'_2, \ldots M'_k \quad . \quad . \quad . \quad . \quad . \quad (14)$$

and the sizes of partial samples, say

$$m'_1, m'_2, \ldots m'_k \quad . \quad . \quad . \quad . \quad . \quad (15)$$

are arbitrary.

In many practical cases the types of sampling described above cannot be applied. Random sampling means the method of including in the sample single elements of the population with equal chances for each element. Human populations are rarely spread in single individuals. Mostly they are grouped. There are certainly exceptions. For instance, when we consider the population of insured persons, they may appear in books of the insurance offices as single units. This circumstance has been used among others by A. B. Hill,* who studied sickness of textile workers, using a random sample of persons insured in certain Approved Societies. But these cases are rather the exceptions. The process of sampling is easier when the population from which we want a sample to be drawn is not a population of persons who are living miles apart, but some population of cards or sheets of paper on which are recorded the data concerning the persons. But even in this simplified position we rarely find ungrouped data. Mostly, for instance when we have to take a sample from the general census data, these are grouped in some way or other, and it is exceedingly difficult to secure an equal chance for each individual to be included in the sample. The grouping of the general census data—for the sake of definiteness we shall bear this example in mind—has generally several grades. The lowest grade consists perhaps in groupings according to lodgings : the inhabitants of one apartment are given a single sheet. The next grouping may include sheets corresponding to apartments in several neighbouring houses † visited by the same officer collecting the data for the Census. These groups are then grouped again and again. Obviously it would be practically impossible to sample at random single individuals from data subject to such complex groupings. Therefore it is useful to consider some further types of the random sampling method.

(*c*) Suppose that the population Π of M' individuals is grouped into M_0 groups. Instead of considering the population Π we may

* A. B. Hill : *Sickness amongst Operatives in Lancashire Cotton Spinning Mills*, London 1930.

† This was the grouping used in the Polish General Census in 1931. The corresponding groups will be called " statistical districts." The number of persons in one statistical district varied from 30 to about 500.

now consider another population, say π, having for its elements the M_0 groups of individuals, into which the population II is divided. Turning to the example of the Polish Census, in which the material has been kept in bundles, containing data from single statistical districts, it was possible to substitute the study of the population π of $M_0 = 123{,}383$ statistical districts, for the study of the population II of $M' = 32$ million individuals. If there are enormous difficulties in sampling individuals at random, these difficulties may be greatly diminished when we adopt groups as the elements of sampling. This being so, it is necessary to consider, whether and how our original problem of estimating X, the average value of the character x of individuals forming the population II, may be transformed into a problem concerning the population π of groups of individuals.

The number we wish to estimate is

$$X = \frac{1}{M'} \sum_{i=1}^{M'} (x_i) \quad . \quad . \quad . \quad . \quad . \quad (16)$$

where x_i means the value of the character x of the i-th individual. Obviously there is no difficulty in grouping the terms of the sum on the right-hand side of the above equation so that each group of terms refers to a certain group of individuals, forming the population π. Suppose that these groups contain respectively

$$v_1, v_2, \ldots v_M \quad . \quad . \quad . \quad . \quad . \quad (17)$$

individuals and that the sums of the x's corresponding to these individuals are

$$u_1, u_2, \ldots u_M \quad . \quad . \quad . \quad . \quad . \quad (18)$$

With this notation we shall have

$$M' = v_1 + v_2 + \ldots + v_M = \Sigma(v) \quad . \quad . \quad (19)$$

$$\sum_{i=1}^{M'} (x_i) = u_1 + u_2 + \ldots u_M = \Sigma(u) \text{ (say)} \quad . \quad (20)$$

The problem of estimating X is now identical with the problem of estimating the character of the population π, namely,

$$\mathrm{X} = \frac{\Sigma(u)}{\Sigma(v)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

We have now to distinguish two different cases: (*a*) the number M' of individuals forming the population II is known, and (*b*) this number is not known.

In the first case the problem of estimating X reduces itself to that of estimating the sum of the u's in the numerator of (21). In the other case we have also to estimate the sum of the v's in the denominator and, what is more, the ratio of the two sums. Owing

to the results of S. Bernstein and of R. C. Geary this may be easily done if the estimates of both the numerator and the denominator in the formula giving X are the best linear estimates. The theorem of S. Bernstein * applies to such estimates, and states that under ordinary conditions of practical work their simultaneous distribution is representable by a normal surface with constants easy to calculate. Of course there is the limiting condition that the size of the sample must be large. The result of Geary † then makes it possible to determine the accuracy of estimation of X by means of the ratio of the separate estimates of the numerator and the denominator.

Thus we see that if it is impossible or difficult to organize a random sampling of the individuals forming the population to be studied, the difficulty may be overcome by sampling groups of individuals. Here again we may distinguish the two methods of unrestricted and of stratified sampling. It is indisputable that the latter has definite advantages both from the point of view of the accuracy of results and of the ease in performing the sampling. Therefore we shall further consider only the method of stratified sampling from the population π, the elements of which are groups of individuals forming the population Π. It is worth noting that this form of the problem is very general. It includes the problem of unrestricted sampling, as this is the special case when the number of strata $k = 1$. It includes also the problem of sampling individuals from the population Π, as an individual may be considered as a group, the size of which is $v = 1$. We shall see further on that the method of stratified sampling by groups includes as a special case the method of purposive selection.

2. *The Method of Purposive Selection.*

Professor Bowley did not consider in his book the above type (*c*) of the method of random sampling by groups.‡ When, therefore, he speaks about the principle of random sampling he is referring to the sampling of individuals. According to Bowley, the method of purposive selection differs from that of random sampling mainly in the circumstances that " in purposive selection the unit is an aggregate, such as a whole district, and the sample is an aggregate of these aggregates, while in random selection the unit is a person or thing, which may or may not possess an attribute, or with which some measurable quantity is associated. . . . Further, the fact that the selection is purposive very generally involves intentional dependence

* S. Bernstein: " Sur l'extension du théorème limite du calcul des probabilités." *Math. Ann.*, Bd. 97.

† R. C. Geary: " The Frequency Distribution of the Quotient of Two Normal Variates." *J.R.S.S.*, Vol. XCIII, Part III.

‡ Though he applied it in practical work.

on correlation, the correlation between the quantity sought and one or more known quantities. Consequently the most important additional investigation in this section relates to the question how far the precision of the measurements is increased by correlation, and how best an inquiry can be arranged to maximize the precision."

It is clear from this quotation that the terminology of Professor Bowley and that which I am using do not quite fit together. In fact the circumstance that the elements of sampling are not human individuals, but groups of these individuals, does not necessarily involve a negation of the randomness of the sampling. Therefore I have thought it useful to consider the special type of random sampling by groups, and the nature of the elements of sampling will not be further considered as constituting any essential difference between random sampling and purposive selection.

The words purposive selection will be used to define the method of procedure described by Bowley, Gini and Galvani. This may be divided into two parts : (*a*) the method of obtaining the sample, and (*b*) the method of estimation of such an average as X, described above.

The method of obtaining the sample assumes that the population Π of individuals is divided into several, M, districts forming the population π, that the number of individuals in each district, say v_i, is known and, moreover, that there is known for each district the value of one or more numerical characters, which Professor Bowley calls " controls." There is no essential difference between cases where the number of controls is one or more, so we shall consider only the case where there is one control, which we shall denote by y_i for the i-th district. We shall retain our previous notation and denote by u_i the sum of values of x, corresponding to the i-th district or group. Consider next, say, $\bar{x}_i = u_i/v_i$ or the mean value of the character x in the i-th district. The basic hypothesis of the method of purposive selection is that the numbers $\bar{x}_i$ are correlated with the control y_i and that the regression of $\bar{x}_i$ on y_i is linear. As we shall have to refer again to this hypothesis, it will be convenient to describe it as the hypothesis H.

Assuming that the hypothesis H is true, the method of forming the sample consists in " purposive selection " of such districts for which the weighted mean

$$Y' = \frac{\Sigma(vy)}{\Sigma(v)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

has the same value, or at least as nearly the same as it is possible, as it has for the whole population, say Y. It is assumed that the above method of selection may supply a fairly representative sample, at least with regard to the character x. As it follows from the quotation from the work of Gini and Galvani, it was also believed that by

multiplying the controls it would be possible to obtain what could be termed a generally representative sample with regard to many characters. Otherwise the method of purposive selection could not be applied to supply a sample which could be used in the future for purposes not originally anticipated.

This is the method of obtaining the sample. As we shall easily see, it is a special case of stratified random sampling by groups. In fact, though the three authors think of districts as of rather large groups with populations attaining sometimes one million persons, they assume that the number M of these districts is not very small. In the Italian investigation it was over 200. If we consider the values of the control, y, calculated for each district, we shall certainly find such districts for which the value of y is practically the same. Thus the districts may be grouped in strata, say of the first order

$$\pi_{y_1}, \pi_{y_2}, \ldots \pi_{y_k} \quad . \quad . \quad . \quad . \quad . \quad (23)$$

each corresponding to a given value of y. Now each of the first order strata of districts may be subdivided into several second order strata, according to the values of v in the districts. Denote by π_{yv} a stratum containing, say, M_{yv} districts, all of which have practically the same values of the control, y, and the same number of individuals v. Denote further by m_{yv} the number of the districts belonging to π_{yv} to be included in the sample. If the principle directing the selection consists only in the fulfilment of the condition that the weighted mean of the control with v's as weights should be the same in the sample and in the population, then it means nothing but a random sampling of some m_{yv} districts from each second order stratum, the numbers m_{yv} being fixed in advance, some of them being probably zero. This is obvious, since for purposes of keeping the weighted mean $Y' = Y =$ constant, two different districts belonging to the same second order stratum are of equal value. Hence we select one of them at random.*

Thus we see that the method of purposive selection consists, (*a*) in dividing the population of districts into second order strata according to values of y and v, and (*b*) in selecting randomly from each stratum a definite number of districts. The numbers of samplings

* It must be emphasized that the above interpretation of the method of purposive selection is a necessary one if we intend to treat it from the point of view of the theory of probability. There is no room for probabilities, for standard errors, etc., where there is no random variation or random sampling. Now if the districts are *selected* according to the corresponding values of the control y and also of the number of individuals, v, they contain, the only possible variate which is left to chance is $\bar{x}_i$. If the districts are very large and therefore only very few, then the majority of second order strata will contain no or only one district. In this case, of course, the process of random sampling from such a stratum is an imaginary one.

are determined by the condition of maintenance of the weighted average of the y. Comparing the method of purposive selection with that of stratified sampling by groups we have to bear in mind these two special features of the former.

IV. Comparison of the two Methods of Sampling.

1. *Estimates of Bowley and of Gini and Galvani.*

Suppose now the sample is drawn and consider the methods of estimation of the average X. In this respect the Italian statisticians do not agree with Bowley, so we shall have to consider two slightly different procedures. I could not exactly follow the method proposed by Professor Bowley. It is more clearly explained by the Italian writers, but I am not certain whether they properly understood the idea of Bowley. It consists in the following :

Denote by X_Σ the weighted mean of values $\bar{x}_i$ deduced from the sample, Σ; by $\bar{x}$, the unweighted mean of the same numbers, also deduced from the sample. Y will denote the weighted mean of the control y, having *ex hypothesi* equal values for the sample and for the population. $\bar{y}$ will denote the unweighted mean of the control y, calculated for the population, and finally g the coefficient of regression of $\bar{x}_i$ on y_i, calculated partly from the sample and partly from the population.

As a first approximation to the unknown X, X_Σ may be used. But it is possible to calculate a correction, K, to be subtracted from X_Σ so that the difference $X_\Sigma - K$ should be considered as the second approximation to X. The correction K is given by the formula

$$K = -(X - \bar{x}) + g(Y - \bar{y}) \quad . \quad . \quad . \quad (24)$$

As the value of X is unknown, its first approximation X_Σ may be substituted in its place. In this way we get as a second approximation to X the expression, say,

$$X' = X_\Sigma + (X_\Sigma - \bar{x}) - g(Y - \bar{y}) \quad . \quad . \quad . \quad (25)$$

I do not know whether this is the method by which Bowley has calculated the very accurate estimates in the examples he considers in his paper. At any rate the method as described above is inconsistent : even if applied to a sample including the whole population and even if the fundamental hypothesis H about the linearity of regression of $\bar{x}_i$ on y_i is exactly satisfied, it may give wrong results :

$$X' \neq X \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (26)$$

This may be shown on the following simple example. Suppose

that the population π consists only of four districts characterized by the values of $\bar{x}_i$, y_i and v_i as shown in the following Table I.

TABLE I.

Districts.	$\bar{x}_i$.	y_i.	v_i.	$u_i = \bar{x}_i v_i$.	$y_i v_i$.
I.	·07	·09	100	7	9
II.	·09	·09	400	36	36
III.	·11	·12	100	11	12
IV.	·13	·12	900	117	108
Totals	·40	·42	1500	171	165
Means	$\bar{x} = ·100$	$\bar{y} = ·105$	—	$X = ·114$	$Y = ·110$

Owing to the fact that the control y has only two different values, ·09 and ·12, there is no question about the hypothesis H concerning the linearity of regression, which is certainly satisfied. The regression line passes through the points with co-ordinates ($y = ·09$, $x = ·08$) and ($y = ·12$, $x = ·12$). Thus the coefficient of regression $g = \frac{4}{3}$. Assume now we have a sample from the above population, which includes the whole of it and calculate the estimate X' of $X = ·114$. We shall have

$$\begin{aligned} X_\Sigma &= &&= ·114 \\ +(X_\Sigma - \bar{x}) &= &&= ·014 \\ -g(Y - \bar{y}) &= -\frac{·02}{3} &&= -·007 \quad . \quad . \quad . \quad (27) \\ & & X' &= ·121, \end{aligned}$$

which is not equal to $X_\Sigma = ·114$.

Gini and Galvani applied Bowley's method to estimate the average rate of natural increase of the population of Italy, using a sample of 29 out of 214 circondarî. They obtained results which they judged to be unsatisfactory, and they proposed another method of estimation. This consists in the following:

They start by finding what could be called the weighted regression equation. If there are several controls, say $y^{(1)}$, $y^{(2)}$, . . . $y^{(s)}$, the weighted regression equation

$$x = b_0 + b_1 y^{(1)} + b_2 y^{(2)} + \ldots + b_s y^{(s)} \quad . \quad . \quad (28)$$

is found by minimizing the sum of squares

$$\Sigma v_i (x_i - b_0 - b_1 y_i^{(1)} - \ldots - b_s y_i^{(s)})^2 \quad . \quad . \quad (29)$$

with regard to the coefficients b_0, b_1, b_2, . . . b_s. This process would follow from the ordinary formulæ if we assumed that one district with the number of individuals v_i and the mean character $\bar{x}_i$ is equivalent to v_i individuals, each having the same value of the character

$x = \bar{x}_i$. Having noticed this, it is not necessary to go any further into the calculations. If there is only one control, y, then the weighted regression equation will be different from the ordinary one in that it will contain weighted sample means of both $\bar{x}_i$ and $\bar{y}_i$ instead of the unweighted ones, and that in the formula of the regression coefficient we should get weighted instead of unweighted sums. The weighted regression equation is then used by Gini and Galvani to estimate the value of $\bar{x}_i$ for each district, whether included in the sample or not. This is done by substituting into the equation the values of the control y_i corresponding to each district and in calculating the value of the dependent variable. The estimates of the means $\bar{x}_i$ thus obtained, say $\bar{x}_i'$, are then used to calculate their weighted mean

$$X' = \frac{\Sigma(v_i \bar{x}_i')}{\Sigma(v_i)}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (30)$$

which is considered as an estimate of the unknown mean X.

Simple mathematical analysis of the situation proved (see Note III) that this estimate is consistent when a special hypothesis, H', about the linearity of regression of $\bar{x}_i$ on y_i holds good, and even that it is the best linear estimate under an additional condition, H_1, concerning the variation of the $\bar{x}_i$ in strata corresponding to different fixed values of y and v.

The hypothesis H' consists in the assumption that the regression of $\bar{x}$ on y is linear not only if we consider the whole population π of the districts, but also if we consider only districts composed of a fixed number of individuals. It is seen that the hypothesis H' is a still more limiting than the hypothesis H.

The other condition, H_1, is as follows. Consider a stratum, π', defined by the values $y = y'$ and $v = v'$ and consider the districts belonging to this stratum. Let

$$\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p \quad . \quad . \quad . \quad . \quad . \quad . \quad (31)$$

be the values of the means $\bar{x}$ corresponding to these districts. The hypothesis, say H_1, under which the estimate of X proposed by Gini and Galvani is the best linear estimate, consists in the assumption that the standard deviation, say σ' of the $\bar{x}_i$ corresponding to the stratum π' may be presented by the formula

$$\sigma' = \frac{\sigma}{\sqrt{v'}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (32)$$

σ being a constant, independent of the fixed value of $v = v'$. This hypothesis would be justifiable if the population of each district could be considered as a random sample of the whole population Π. In fact, then the standard deviation of means, $\bar{x}_i$ corresponding to

districts having their population equal to v would be proportional to $v^{-\frac{1}{2}}$. The population of a single district is certainly not a random sample from the population of the country, so the estimate of Gini and Galvani is not the best linear estimate—at least in most cases.

Having got so far we may consider whether and to what extent there is justification for the principle of choosing the sample so that the weighted mean of the control in the sample should be equal to the weighted mean of the population. The proper criterion to use in judging seems to be the standard error of the estimate of X'. This is given by a function (see Note III) which, *cæteris paribus*, has smaller values when the weighted sample mean of the control is equal to its population value, and when the sum of weights $\Sigma(v)$, calculated for the sample, has the greatest possible value. Thus the principle of purposive selection is justified. The analysis carried out in Note III suggests also that if the number of districts to be included in the sample is fixed we should get greater accuracy by choosing larger districts rather than smaller ones. This conclusion, however, depends largely upon the assumptions made concerning the standard deviations within the districts and the linearity of regression.

2. *The Hypotheses underlying both Methods and the Conditions of Practical Work.*

We may now consider the questions: (1) Are we likely to find in practice instances where the hypotheses underlying the method of purposive selection are satisfied, namely, the hypothesis H' concerning the linearity of regression and the hypothesis H_1 concerning the variation of the character sought within the strata of second order? (2) If we find instances where these hypotheses are not satisfied exactly, then what would be the result of our ignoring this fact and applying the method of purposive selection? (3) Is it possible to get any better method than that of purposive selection? *

With regard to (1), I have no doubt that it is possible to find instances, when the regression of a certain character $\bar{x}_i$ on the control y_i is fairly nearly linear. This may be the case especially when one of the characters $\bar{x}$ and y is some linear function of the other, say if $\bar{x}$ is the rate of natural increase of the population and y the birth-rate. This is the example considered by Gini and Galvani. I think, however, that this example is rather artificial. When y is known for any district, in most cases we shall probably have all the necessary data to enable us to compute the $\bar{x}$ without any appeal to the representative method. In other cases, however, when the connection between the character sought and the possible control is not so straightforward, I think it is rather dangerous to assume

* *I.e.* a method which would not lose its property of being consistent when the hypothesis H' is not satisfied.

any definite hypothesis concerning the shape of the regression line. I have worked out the regression of the mean income $\bar{x}_i$ of people inhabiting different circondarî on the first of the controls used by Gini and Galvani, *i.e.* the birth-rate, y_i. The figures I and II give respectively the approximate spot diagram of the correlation table of those characters, and the graph of the weighted regression line of $\bar{x}_i$ and y_i. It is to be remembered that the data concern the whole population, and thus the graph represents the " true " regression line. This is far from being straight. It is difficult, of course, to judge how often we shall meet in practice considerable divergencies from linearity. I think, however, that it is rather safer to assume that the linearity is not present in general and to consider the position when the hypothesis H' is not satisfied.

The hypothesis H_1 is probably never satisfied.

With regard to (2) : Note III shows that the estimate of Gini and Galvani generally ceases to be unbiased when we can no longer make any assumption about the shape of the regression line of $\bar{x}$ on y. It may be kept consistent only by adjusting in a very special manner the numbers of districts selected from single second order strata. In fact the consistency requires that the number of districts, say m' to be selected from a stratum containing altogether M' districts, should satisfy the condition

$$\frac{m'}{M'} = \frac{\Sigma(v) \text{ for the sample}}{\Sigma(v) \text{ for the population}} \quad . \quad . \quad . \quad (33)$$

Any departure from this rule may introduce some bias in the estimate.

With regard to (3) : There is no essential difficulty in applying Markow's method to find the best unbiased estimates of the average X determined from a sample obtained by the method of stratified sampling by groups. This has been done in full detail in my Polish publication (there is an English summary) * concerning the theory of the representative method. The principle of stratifying, *i.e.* of the division of the original population of districts into strata, does not affect the method of obtaining the estimate. In any case, and whatever the variances of the $\bar{x}_i$ within the strata, the best linear estimate of X is always the same.

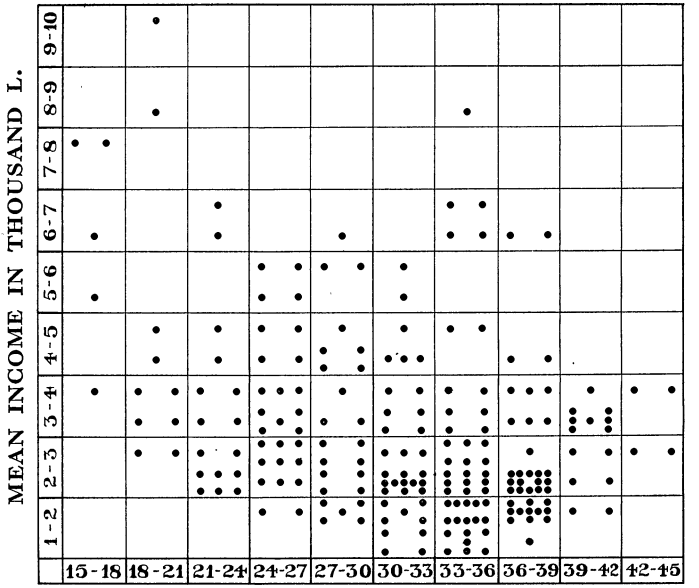
I shall return here to variables introduced previously and shall use

$$u_i = v_i \bar{x}_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (34)$$

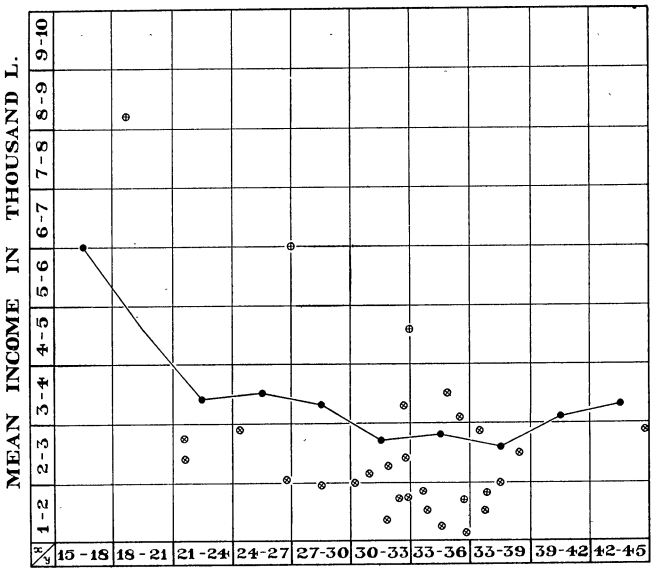
instead of $\bar{x}_i$. Suppose that in m' samplings from a stratum containing M' districts, we obtained m' different values of u. Denote by $\bar{u}$ their arithmetic mean. Then the product $M'\bar{u}$ will be the

* J. Neyman : *An Outline of the Theory and Practice of Representative Method, Applied in Social Research.* Institute for Social Problems, Warsaw, 1933.

FIG.1.



REGRESSION OF x ON y.



SPOTS CORRESPOND TO DISTRICTS FORMING THE SAMPLE.

FIG.2.

estimate of the sum of the u's for the whole stratum. Summing these estimates for all strata, we get the best estimate of the sum of u's for the whole population. To get an estimate of X it remains only to divide the estimate of the sum of u's by the sum of v's, which may be known or may be estimated by the same method. Thus the final estimate of X say X'' is either

$$X'' = \frac{\Sigma(M'\bar{u})}{\Sigma(v)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (35)$$

if the v's are known for every district, or in the other case

$$X'' = \frac{\Sigma(M'\bar{u})}{\Sigma(M'\bar{v})} \quad . \quad . \quad . \quad . \quad . \quad . \quad (36)$$

where $\bar{v}$ means the arithmetic mean of v's, calculated from the sample separately for each stratum.

The consistency of the estimates $\Sigma(M'\bar{u})$ and $\Sigma(M'\bar{v})$ does not depend upon any arbitrary hypothesis concerning the sampled population. The only condition, which must be satisfied is that the sample should contain districts from every stratum. So we may safely apply these estimates, whatever the properties of single strata and irrespective of variations of u's and v's within the strata. But the standard errors of the two estimates do depend both upon the variability of the characters of districts within the strata and upon the relationship of numbers m' and M'. It is known that the formula giving the variance, say σ^2, of the estimate $\Sigma(M'\bar{u})$ is as follows :

$$\sigma^2 = \sum\left\{\frac{M_i^2}{m_i}\frac{M_i - m_i}{M_i - 1}\sigma_i^2\right\} \quad . \quad . \quad . \quad . \quad (37)$$

where m_i and M_i refer to the i-th stratum, σ_i^2 is the variance of the u's in the i-th stratum and the summation Σ extends over all strata. The dependence of σ^2 upon the σ_i^2 is obvious. If we succeed in dividing the population π into strata which would be very homogeneous with regard to the character u of the districts, σ_i^2 will be small and so will be σ^2. It is also obvious that by increasing the numbers, m_i, of districts to be selected from the strata we shall also improve the accuracy of the estimate. By taking $m_i = M_i$ the accuracy will be absolute, but then we shall have an exhaustive enquiry. It will probably be necessary to assume that the actual conditions of the research fix a certain number, say

$$m_0 = \Sigma(m_i) \quad . \quad . \quad . \quad . \quad . \quad . \quad (38)$$

of districts to be selected from the population. Our problem will then consist in distributing the total number of samplings among single strata so as to have the minimum possible value of σ^2.

Simple calculations show that the variance (37) may be written in the form

$$\sigma^2 = \frac{M_0 - m_0}{m_0}\Sigma(M_i S_i^2) + \Sigma m_i\left(\frac{M_i S_i}{m_i} - \frac{\Sigma(M_i S_i)}{m_0}\right)^2 - \frac{M_0}{m_0}\Sigma M_i\left(S_i - \frac{\Sigma(M_i S_i)}{M_0}\right)^2 \quad . \quad . \quad (39)$$

where S_i^2 stands for $M_i\sigma_i^2/(M_i - 1)$. We see that only the middle term of the right-hand side depends upon the values of the m's. The other terms remain constant whatever the system of m's, provided their sum, m_0, remains unchanged. Thus the method of diminishing the value of σ^2 consists in diminishing the middle term of the right-hand side of (39). This has its minimum value, zero, when the numbers m_i are proportional to the products $M_i S_i$. Thus if it is possible to estimate the variances σ_i^2 of the u's within any given stratum, the most favourable system of m_i's is not that for which the m_i are proportional to the M_i. Denote the three terms of the right-hand side of (39) respectively by A, B and $-C$. If we assume that the m_i's are proportional to M_i, then we shall find that the term $B = C$ and the variance σ^2 is reduced to

$$\sigma^2 = \frac{M_0 - M_0}{m_0}\Sigma(M_i S_i^2) = A \quad . \quad . \quad (40)$$

If, however, m_i are proportional to $M_i S_i$, then the positive term B in (39) vanishes and we get

$$\sigma^2 = A - C \quad . \quad . \quad . \quad . \quad . \quad . \quad (41)$$

which is the optimum value of σ^2.

If the research is carried out with regard to several highly correlated characters of groups forming the elements of sampling, then by means of a preliminary enquiry it is possible to estimate the numbers S_i, which, if calculated for the different characters sought, would be also correlated. Hence we could then by a proper choice of the numbers m_i if not reduce the middle term of the right-hand side of (39) to zero, then at least diminish it sensibly.

Such was the case in the Warsaw enquiry already referred to, carried out by the Institute for Social Problems. The purpose of this enquiry was to describe the structure of the working class in Poland, according to different characters, such as the age distribution of males and females, whether married or single, the distribution of the number of children in families, etc., and this separately for three different categories of workers. Obviously all characters of the elements of sampling sought are highly correlated with the number of workers in each element. As there are in Poland large

districts where the percentage of workers is negligible and others where they are numerous, the numbers S_i calculated for the different characters sought varied from stratum to stratum in broad limits. Accordingly, an adjustment of numbers m_i was made in order to diminish variances of the estimates.

The necessity of these adjustments is not difficult to appreciate. One feels intuitively that it would be unreasonable to include in the sample, equal percentages of statistical districts from two strata A and B in one of which, A, the percentage of workers, amounts to say 60 per cent. and in the other, B, to 5 per cent. It may even be assumed that in such cases it would be advisable to omit totally the stratum B. However, I do not think it is really always advisable, since the total number of workers in the stratum B may be sometimes equal to or even larger than those in stratum A, and the structure of family conditions in both strata may be very different.

Of course this sort of research is a rather special one. In many cases the characters sought are not likely to be highly correlated. In other cases—as in the work of Gini and Galvani—it is impossible to state at the time of sampling which characters of the elements of sampling will be the matter of research. Any adjustments of the numbers, m_i, are then impossible, since a wrong adjustment may give to σ^2 a value larger than that corresponding to the system of proportional sampling. The best we can do is to sample proportionately to the sizes of strata.*

Thus the principle that the numbers m_i should be proportional to M_i, suggested by Professor Bowley, is just the best that one could advise in the most general case.

Up to this point I have considered the possibility of reducing the value of σ^2 by adjusting properly the numbers m_i of samplings from different strata. I assumed, in fact, that the districts forming the elements of sampling and their total number m_0 to be included in the sample are fixed. Now I shall suppose that the districts are not fixed except that their size will not be very different, and that all that is known is that the sample should include a certain percentage of districts, whatever be their kind.

In other words, I intend to consider the situation in which we decide to include in the sample some, *e.g.* 10 per cent. of the population, and are considering the question what should be our "districts," forming the elements of sampling: whether they should include about, say 200 or about 20,000 persons, etc.

I wish to call attention to the fact, that the ratios m_i/M_i being fixed in some way or other, the value of σ^2 (see (37)) depends upon

* It is to be remembered that "the size of the stratum" is the number, M_i, of its elements, not the number of individuals.

the products $M_i S_i^2 = M_i^2 \sigma_i^2/(M_i - 1)$, or practically upon the products $M_i \sigma_i^2$, and may be influenced by a proper choice of the element of sampling. In fact, if we consider two different systems of division of a stratum into larger and smaller districts, then the values of u's corresponding to several smaller districts forming a larger one, will be very generally positively correlated. As the result of this the value of $M_i \sigma_i^2$, corresponding to a subdivision of strata into smaller districts, will be less than that corresponding to a subdivision into larger districts. This point may be illustrated on an extreme case. Suppose, for instance, that X represents the proportion of agricultural workers aged 20 to 21. Then for every individual of the population x will have the value $x = 1$ if this individual is an agricultural worker aged 20 to 21, and $x = 0$ in all other cases. If now we consider as elements of sampling the statistical districts including 50 inhabitants, then in a stratum we may have (in the most unfavourable case) one half of the districts composed only of agricultural workers at the fixed age, thus having $u = 50$, while in the other half of the district $u = 0$. The standard deviation σ_i would be 25. On the other hand, if the districts were to include not 50 persons, but, say, 500, the maximum possible value of σ_i would be tenfold, 250. The term $M_i \sigma_i^2$ in this second case would be ten times larger than in the former. Of course it may be argued that taking larger districts we decrease the chance of their being extremely differentiated. This is certainly so, but on the other hand I think it extremely probable that the products $M_i S_i^2$ calculated for districts including tens of thousands or hundreds of thousands of people must be expected to be incomparably larger than those calculated for the districts including on the average two or three hundred people. And this for the majority of imaginable characters which could be the matter of statistical research.*

The effect of choosing smaller units of sampling may be roughly illustrated on another example of a game of chance, in which the probability of a gain is equal to $\frac{1}{2}$. Suppose we dispose of a sum of £100 for the game, which we may either bet at once or divide in a hundred separate bettings. In the first case it is obviously impossible to predict the result. In the other case, however, we may

* I do not know whether these were the reasons for which Gini and Galvani expressed the view that the results of their sampling would have been much better if the method of selection adopted were that of stratified sampling, and if the element of sampling were a commune. The reasons for not applying this method seems to be that "nobody could under-appreciate the difficulty in a stratification of the communes simultaneously with regard to different characters." (Page 6, *loc. cit.*). I think, however, that a stratification assuming the 214 circondarî as strata, each containing about 40 communes, which might be considered as elements of sampling, would be quite sufficient. Of course the results would be probably still better if the elements of sampling were smaller than a commune.

be pretty certain that the gain or loss will not exceed some £15 or £20.

Similarly, if we want to obtain a representative sample, say amounting to 15 per cent. of the population, it is much safer to make, say, 3,000 samplings of small units rather than 30 of larger ones, and this is probably true, whatever the stratification.

3. *Numerical Illustration.*

It may be perhaps useful to consider a simple numerical example showing the effect on the accuracy of the method of purposive selection of non-linearity of regression of the character sought on the control.

We shall consider the result of sampling from four populations, in one of which the weighted regression of $\bar{x}$ on y is linear, and in three others where it is showing different degrees of deviation from linearity. All four populations are divided into three strata according to the values of the control $y = -1$, $y = 0$ and $y = +1$. Each stratum contains three districts. The construction of the population, say π_1 with linear weighted regression is shown in Table II.

TABLE II.

$y = -1$.			$y = 0$.			$y = +1$.		
No. of District i.	u_i.	v_i.	No. of District i.	u_i.	v_i.	No. of District i.	u_i.	v_i.
1	−17	1	4	1	3	7	20	3
2	−18	2	5	0	2	8	18	2
3	−19	3	6	−1	1	9	16	1
Totals	−54	6	—	0	6	—	54	6
Means	$\bar{x}(-1) = -9$	—	—	$\bar{x}(0) = 0$	—	—	$\bar{x}(1) = 9$	—

As in the actual calculations we have to use the products $\bar{x}_i v_i = u_i$ I have omitted the values of the $\bar{x}$'s and have given the values of the u's instead. It is easy to see that the population values $X_1 = Y = 0$. The weighted averages of the $\bar{x}$'s in each array are given at the bottom, namely − 9, 0, + 9, and it is seen that the regression is linear.

The populations π_2, π_3 and π_4 may be obtained from the population π_1 so easily that it is not necessary to describe them in special tables. The population π_2 is obtained by keeping the strata corresponding to $y = -1$ and $y = +1$ unchanged and by adding to each value of u_i in the stratum $y = 0$ the same number, 6. As a result of this the weighted mean of $\bar{x}_i$, say $\bar{x}(0)$ in the middle stratum will be raised to $\bar{x}(0) = 3$ and the regression will cease to be linear.

X will now have the value $X_2 = 1$. The population π_3 will be obtained from the population π_2 in the same way as this was obtained from the population π_1. Similarly, the population π_4 will be obtained from π_3 by the same operation. The values of the weighted mean of x's in the stratum $y = 0$ and in the populations will be as follows :

$$\bar{x}_3(0) = 6, \quad X_3 = 2, \quad . \quad . \quad . \quad . \quad . \quad (42)$$
$$\bar{x}_4(0) = 9, \quad X_4 = 3.$$

I then considered all possible samples from these populations, subject to the conditions : (*a*) $\Sigma(v) = 7$, *i.e.* the number of individuals in the sample (not the number of elements of the sample) is fixed in advance, and (*b*) the sample weighted mean of the control y should be equal to its population value $Y = 0$. The details of the results obtained are given in the following Table III :

TABLE III.

Populations.	π_1.	π_2.	π_3.	π_4.	All popul.
Districts.	$\Delta' = X'$.	Δ'.	Δ'.	Δ'.	Δ''.
1, 2, 6, 7	−2·29	−2·43	−2·57	−2·71	·25
1, 2, 6, 8, 9	− ·29	− ·43	− ·57	− ·71	−·25
3, 6, 7	·00	− ·14	− ·29	− ·43	·00
3, 6, 8, 9	2·00	1·86	1·71	1·57	−·50
2, 4, 8	·14	·00	− ·14	− ·28	·17
2, 5, 6, 8	− ·14	·57	1·29	2·00	−·08
1, 4, 5, 9	·00	·71	1·43	2·14	−·08

Here X' and X'' mean the estimates of X, (i) obtained by method proposed by Gini and Galvani, and (ii) calculated from the formula (35). $\Delta' = X' - X$ and $\Delta'' = X'' - X$ represent the errors of these estimates. It will be seen that the estimate X'' gives generally better results. But this is not an essential point in the example, as it is easy to construct another in which the estimate X' would be the better. In fact, the accuracy of X'' is connected with the variability of the u's within the strata. If in single strata corresponding to different values of y, the variation of the u's is very large, then the results obtained by using X'' would not be very good. The comparison between two methods could perhaps be worked out arithmetically if we were to consider second order strata. But this would extend the example to the point of losing its illustrative properties.

What is important to note is that the results obtained by using X' get worse and worse with the departure from the linearity of regression. This last circumstance does not affect the accuracy of X'' at all. On the other hand, a change in the values of σ_i would affect X''.

V. Conclusions.

Let us now turn to the question, which I raised at the beginning of the paper, whether the idea of a certain equivalency of the two aspects of the representative method is really justified. We shall have to consider both the theory and the practical results obtained by both methods. Professor Bowley, who was first to give the theory of the method of purposive selection, has not, I believe, used it in practice. The most important research, known to me, by which the representative method was used, is the *New Survey of London Life and Labour*. It has been directed by Bowley, who chose the method of random sampling by groups. This is, I think, an example of the intuition to which Laplace referred.

The Italian statisticians, who applied the method of purposive selection of very few (29) and very large districts with populations from about 30,000 to about 1 million persons, did not find their results to be satisfactory. The comparison between the sample and the whole country showed, in fact, that though the average values of seven controls used are in a satisfactory agreement, the agreement of average values of other characters, which were not used as controls, is often poor. The agreement of other statistics besides the means, such as the frequency distributions, etc., is still worse. This applies also to the characters used as controls. The statement of the above facts is followed in the paper by Gini and Galvani by general considerations concerning the concept of a representative sample. They question whether it is possible to give any precise sense to the words " a generally representative sample." I think it is, and I agree also that an exhaustive enquiry is the only method which can give absolutely true results. However, the need for a representative method is an urgent one and many enquiries would be impossible if we were not able to use this method. In fact we are often forced to apply sampling for general purposes, so as to get a " generally representative " sample, which might be used for a variety of different purposes.

If there are difficulties in defining the " generally representative sample," I think it is possible to define what should be termed a *representative method of sampling* and a *consistent method of estimation*. These I think may be defined accurately as follows. I should use these words with regard to the method of sampling and to the method of estimation, if they make possible an estimate of the accuracy of the results obtained in the sense of the new form of the problem of estimation, *irrespectively of the unknown properties of the population* studied. Thus, if we are interested in a collective character X of a population π and use methods of sampling and of estimation, allowing

us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in the statements

$$X_1(\Sigma) \leq X \leq X_2(\Sigma) \quad . \quad . \quad . \quad . \quad . \quad (43)$$

does not exceed the limit $1 - \varepsilon$ prescribed in advance, *whatever the unknown properties of the population*, I should call the method of sampling representative and the method of estimation consistent. We have seen that the method of random sampling allows a consistent estimate of the average X whatever the properties of the population. Chocsing properly the elements of sampling we may deal with large samples, for which the frequency distribution of the best linear estimates is practically normal, and there are no difficulties in calculating the confidence intervals. Thus the method of random stratified sampling may be called a representative method in the sense of the word I am using. This, of course, does not mean that we shall always get correct results when using this method. On the contrary, erroneous judgments of the form (43) must happen, but it is known how often they will happen in the long run : their probability is equal to ε.

On the other hand, the consistency of the estimate suggested by Gini and Galvani, based upon a purposely selected sample, depends upon hypotheses which it is impossible to test except by an extensive enquiry.

If these hypotheses are not satisfied, which I think is a rather general case, we are not able to appreciate the accuracy of the results obtained. Thus this is not what I should call a representative method. Of course it may give sometimes perfect results, but these will be due rather to the uncontrollable intuition of the investigator and good luck than to the method itself. Even if the underlying hypotheses are satisfied, we have to remember that the elements of sampling which it is possible to use when applying the purposive selective method, must be very few in number and very large in size. Consequently I think that when using this method we are very much in the position of a gambler, betting at one time £100.

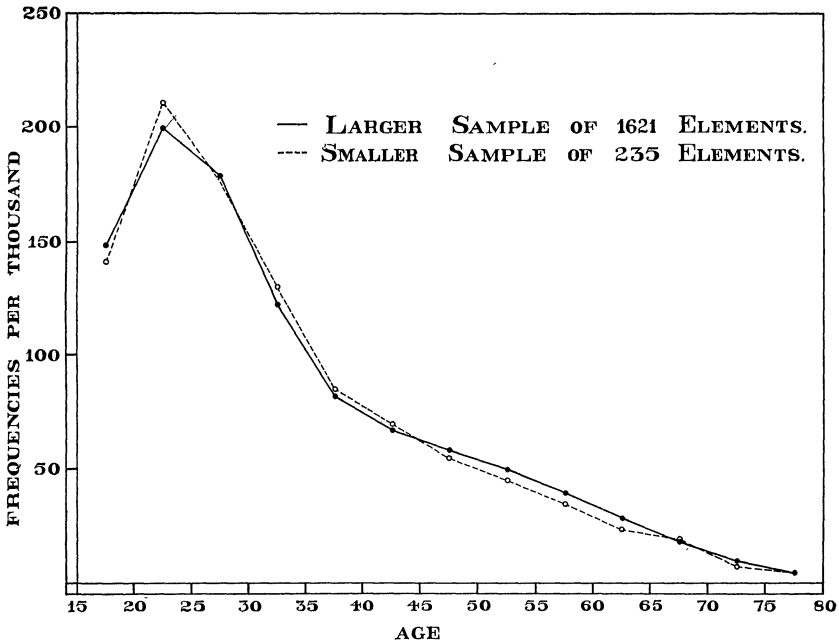
For the above reasons I have advised the Polish Institute for Social Problems to use the method of random stratified sampling by groups when carrying out the enquiry on the structure of Polish workers.*

Poland was divided into 113 strata, containing 123,383 elements of sampling (statistical districts). The average number of persons within an element of sampling was about 250 persons. There were

* The results of this enquiry are to be found in the publication of J. Piekalkiewicz : *Rapport sur les recherches concernant la structure de la population ouvrière en Pologne selon la méthode répresentative.* Institute for Social Problems, Warsaw, 1934.

considerable variations from stratum to stratum. The random stratified sample contained altogether 1,621 elements, thus about 1·24 per cent. of the whole population. I am not yet able to state how accurate are the results obtained, as the respective data of the General Census are not yet published. All that was possible in testing their accuracy was to compare the age distribution of workers found in the whole sample with the age distribution computed from a minor sample of 235 elements selected for an introductory enquiry

FIG III.



which aimed at testing the variability within the strata. The results are presented in Figure III and in Table IV, and seem to be satisfactory. However, even if through the chances of sampling they had been bad, I think I was justified in advising the method of stratified sampling by groups, because I was able to calculate that (with the probability of an error equal to ·01) the error of actuarial calculations, based upon the tables which were computed as the result of enquiry, could not exceed 4·5 per cent.

The method of stratified sampling by groups has been recently used by Professor O. Anderson,* who directed an enquiry into the farm-

* *Bull. de Statistique*, publ. Direction Gen. de Statistique de Bulgarie, No. 8, 1934.

ing conditions in Bulgaria. The process of getting the sample with which he was faced was a more difficult one, as this was not a sample of sheets of paper containing the necessary information, but a sample of villages from which it was necessary to collect the original data. In fact the enquiry in question was a substitute for a general agricultural census. The element of the sampling was a village. The total number of about 5,000 villages was divided into 28 strata. Out of each stratum 2 per cent. of the villages were selected to form the sample. There is only one detail in this enquiry which I am not certain is justifiable. When selecting the villages from single strata special attention was paid to selecting villages which according to the last General Census in 1926 showed a dis-

TABLE IV.

Age Distribution of Polish Workers.

Males.

Age.	Larger Sample.	Smaller Sample.
15–19	148	141
20–24	199	213
25–29	178	176
30–34	122	130
35–39	82	85
40–44	67	69
45–49	58	54
50–54	49	44
55–59	39	34
60–64	28	23
65–69	18	19
70–74	9	7
75–79	4	4
Totals	1001	999

tribution of different characters of farms, similar to that in the whole stratum. I think that the variability of farms and villages is also a character of their population which may be of interest. This character, however, if the efforts of Bulgarian investigators were successful, would be biased in the sample.

The final conclusion which both the theoretical considerations and the above examples suggest is that the only method which can be advised for general use is the method of stratified random sampling. If the conditions of the practical work allow, then the elements of the sampling should be individuals. Otherwise we may sample groups, which, however, should be as small as possible. The examples of enquiries in London, in Bulgaria, and in Poland show that random sampling by groups does not present unsurmountable difficulties.

There are instances when we may select individuals purposely with great success. Such is, for instance, the case when we are interested in regression of some variate y on x, in which case the selection of individuals with values of x varying within broad limits would give us more precision. But these cases are rather exceptional.*

VI. Appendix.

Note I.

Suppose we are taking samples, Σ, from some population π. We are interested in a certain collective character of this population, say θ. Denote by x a collective character of the sample Σ and suppose that we have been able to deduce its frequency distribution, say $p(x|\theta)$, in repeated samples and that this is dependent on the unknown collective character, θ, of the population π.

The collective characters I am speaking about are arbitrary. The position may be illustrated, for instance, by supposing that the collective character θ is the proportion of a certain type of individuals in the population π, and x the proportion of the same type of individuals in the sample. The distribution of x is then a binomial, depending upon the value of θ.

Denote now by $\varphi(\theta)$ the unknown probability distribution *a priori* of θ. Suppose that the general conditions of sampling and the properties of the collective characters θ and x define certain values which these characters may possess. In the example I mentioned above, θ, the proportion of individuals of the given type in the population may be any number between 0 and 1. On the other hand, x, the proportion of these individuals in the sample, say of n, could have values of the form k/n, k being an integer $0 \leq k \leq n$.

The new form of the problem of estimation of the collective character θ may be stated as follows: given any positive number $\varepsilon < 1$, to associate with any possible value of x an interval

$$\theta_1(x) < \theta_2(x) \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

such that if we accept the rule of stating that the unknown value of the collective character θ is contained within the limits

$$\theta_1(x') \leq \theta \leq \theta_2(x'). \quad . \quad . \quad . \quad . \quad . \quad (2)$$

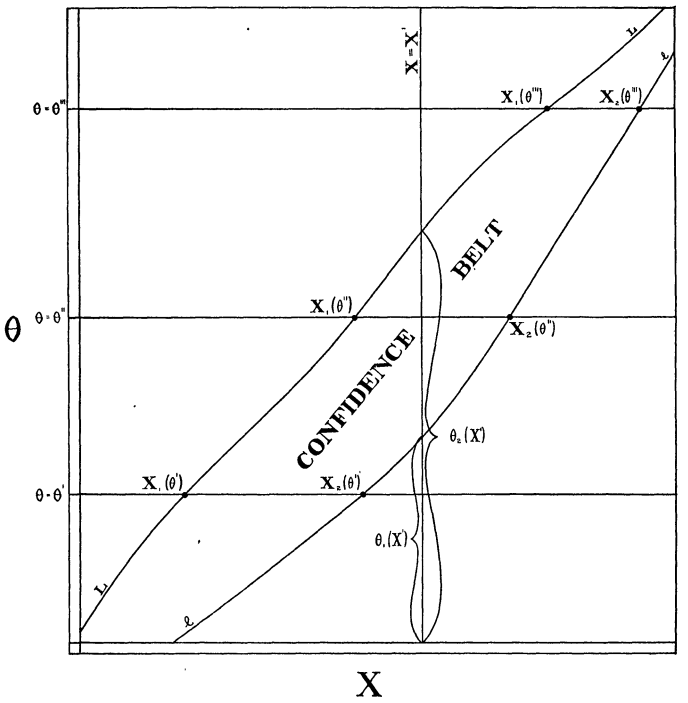
every time the actual sampling provides us with the value $x = x'$, the probability of our being wrong is less than or at most equal to $1 - \varepsilon$, and this whatever the probability law *a priori*, $\varphi(\theta)$.

The value of ε, chosen in a quite arbitrary manner, I propose

* Interesting remarks in this respect are to be found in the excellent book of M. Ezekiel: *Methods of Correlation Analysis* (1930).

to call the " confidence coefficient." If we choose, for instance, $\varepsilon = \cdot 99$ and find for every possible x the intervals $[\theta_1(x), \theta_2(x)]$ having the properties defined, we could roughly describe the position by saying that we have 99 per cent. confidence in the fact that θ is contained between $\theta_1(x)$ and $\theta_2(x)$. The numbers $\theta_1(x)$ and $\theta_2(x)$ are what R. A. Fisher calls the fiducial limits of θ. Since the word " fiducial " has been associated with the concept of " fiducial pro-

FIG. IV.



bability " which has caused the misunderstandings I have already referred to, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals, corresponding to the confidence coefficient ε. The solution of the problem thus stated is an immediate one.

Consider the plane on which the rectangular axes of coordinates OX and $O\theta$ are drawn. Fix any possible value of θ, say $\theta = \theta'$, and find on the straight line

$$\theta = \theta' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

an interval, say $x_1(\theta')$ to $x_2(\theta')$, having the property that in cases when θ' is the true value of the collective character θ, the probability, say $P(\theta')$, of having from the sample x within the limits

$$x_1(\theta') \leq x \leq x_2(\theta') \quad . \quad . \quad . \quad . \quad . \quad (4)$$

is larger than or at least equal to ε. Obviously the limits $x_1(\theta')$ and $x_2(\theta')$ may be always found, and this generally in many different ways. If the variate x is a continuous one, then the limits $x_1(\theta')$ and $x_2(\theta')$ may be fixed so as to have rigorously

$$P(\theta') = \varepsilon \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

In the case, however, when the variate x is not continuous, for instance if it follows the binomial law of frequency, we should have only

$$P(\theta) > \varepsilon \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

For the sake of definiteness, we shall assume that the interval $[x_1(\theta'), x_2(\theta')]$ is chosen to be the shortest possible satisfying the condition (6). Such an interval will be called the interval of acceptance, corresponding to the chosen value of ε. Suppose now we have found the intervals of acceptance, corresponding to all possible values of θ. Now join all left-hand side boundaries of the intervals of acceptance by a continuous line, which may be a smooth curve or a polygon. Denote this line by *LL*. Another line, say *ll*, will join the right-hand side boundaries of the intervals of acceptance. The two lines *LL* and *ll* will be boundaries of a certain belt, which I shall call the confidence belt, *CB*.

Consider now the points, say *A* with coordinates $(x_1\theta)$, thus representing combinations of all possible values of x and θ. The confidence belt *CB* as defined above has the fundamental property that whatever the probability law *a priori* $\varphi(\theta)$, the probability of having the point *A* inside of the confidence belt is equal to or larger than the chosen value of the confidence coefficient ε. This probability may be represented either by means of integrals or by means of the sums extending over all possible positions of the point *A* inside *CB*, according to the properties of the variates x and θ, which may be continuous or not. Thus if p_{CB} is the probability under consideration, we should have either the expression

$$p_{CB} = \int\int_{CB} \varphi(\theta) p(x|\theta) dx d\theta, \quad . \quad . \quad . \quad . \quad (7)$$

or

$$p_{CB} = \sum_{\theta} \varphi(\theta) \sum_{x_1(\theta) \leq x}^{x \leq X_2(\theta)} p(x|\theta), \quad . \quad . \quad . \quad . \quad . \quad (8)$$

or

$$p_{CB} = \sum_{\theta} \varphi(\theta) \int_{x_1(\theta)}^{x_2(\theta)} p(x|\theta) dx, \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

or finally

$$p_{CB} = \int \varphi(\theta) \sum_{x_1(\theta)}^{x_2(\theta)} p(x|\theta) d\theta, \quad . \quad . \quad . \quad . \quad . \quad (10)$$

all summations and integrations extending over the area of CB. It will be noticed that the case of θ and x being proportions of certain individuals in the population and in the sample could lead us either to the formula (8) or to the formula (10) according to the assumptions about the population sampled, which may be finite or infinite.

Whatever the case may be and whichever of the four formulæ may accurately correspond to the actual conditions of the problem, the summations and the integrations may be executed in the same order. We sum first (or integrate) for different values of x, that is for the values contained in the interval of acceptance, corresponding to a fixed value of θ. This gives us the probability $P(\theta)$, which owing to the special choice of the interval of acceptance is $\geqslant \varepsilon$. Substituting ε in the formula giving p_{CB} for the integral or sum with regard to x, we get either

$$p_{CB} = \int \varphi(\theta) P(\theta) d\theta \geqslant \varepsilon \int \varphi(\theta) d\theta = \varepsilon \, . \quad . \quad . \quad . \quad (11)$$

or

$$p_{CB} = \sum_{\theta} \varphi(\theta) P(\theta) \geqslant \varepsilon \sum_{\theta} \varphi(\theta) = \varepsilon \quad . \quad . \quad . \quad . \quad . \quad (12)$$

where the integration or summation with regard to θ extends over all possible value of this character. Thus the respective integral or sum of $\varphi(\theta)$ is equal to one. The above formulæ complete the proof of the fundamental property of the confidence belt. It is to be noted that if x is continuous and the intervals of acceptance are so chosen that $P(\theta) = \varepsilon$ for every θ, then, whatever the unknown distribution *a priori* of θ, we should have

$$p_{CB} = \varepsilon \, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

The construction of the confidence belt is quite independent of any arbitrary assumption concerning the values of θ. If the confidence belt is constructed, we may affirm that the point A *will* lie inside of the belt. This statement may be erroneous, but the probability of the error is either equal to or less than $1 - \varepsilon$—thus is as small as desired.

The solution of the problem of estimation consists in constructing the confidence belt and in affirming that the point A, representing the combination of some possible value of x with some possible value of θ, will lie inside of the belt. When observation provides

us with the value of $x = x'$, we shall consider this as additional, and this time accurate, information about the position of the point A, and shall combine it with the previous (uncertain) statement that A lies inside CB. We shall draw the line parallel to the axis of θ and corresponding to the equation

$$x = x' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

and find its intersections with the boundaries ll and LL of the confidence belt. Denote by $\theta_1(x')$ and $\theta_2(x')$ the ordinates of these points of intersections. The interval $\theta_1(x')$, $\theta_2(x')$ will be the confidence interval corresponding to $x = x'$. Stating that

$$\theta_1(x') \leqslant \theta \leqslant \theta_2(x') \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

every time the observation gives us $x = x'$ we may be wrong. This will happen only if the point A happens to be outside the confidence belt, but the probability of this last fact is equal to $1 - \varepsilon$.*

Note II. The Markoff Method and Markoff Theorem on Least Squares.

The importance of the work of Markoff concerning the best linear estimates consists, I think, chiefly in a clear statement of the problem. The subsequent theory is a matter of easy algebra.

Suppose we have a sample of n values

$$x_1, x_2, \ldots x_n \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

each being randomly drawn from some population $\pi_i (i = 1, 2, \ldots n)$. Denote by A_i and σ_i the mean and the standard deviation in the population π_i. Suppose further that it is known that

$$A_i = a_{i1}p_1 + a_{i2}p_2 + \ldots + a_{is}p_s \quad . \quad . \quad . \quad (2)$$

where the p's are some unknown parameters, and the a's known coefficients and where $s \leqslant n$. Consider now the problem of finding

* The theory developed by Fisher runs on somewhat different lines. It applies only to the case just described when we know the distribution of the collective character x depending upon only one unknown character. The method I am using seems to have the advantage that it allows an easy generalization to the case where there are many unknown parameters in the frequency distribution of several variates describing the results of sampling, while we are interested only in the value of some of them. My method of approach seems to have also the advantage that starting with the calculations depending explicitly upon the unknown probability law *a priori* it shows exactly how this dependence is being eliminated. This theory has been partly set out in my lectures at the University of Warsaw and is now the chief topic of my lectures delivered at University College. It is hoped it will be soon published as an issue of a lecture series delivered at the Department of Applied Statistics at University College, London. The case when θ and x are proportions of some individuals in the population and in the sample is discussed in the paper by C. J. Cloper and E. S. Pearson (now in print) giving some interesting remarks on the relationship between the concepts of confidence intervals and the probabilities *a posteriori* in the sense of Bayes. This paper contains also graphs with a system of confidence belts, corresponding to different sizes of the samples and to different confidence coefficients.

the best linear estimate of a collective character of the populations defined by the equation

$$\theta = b_1 p_1 + b_2 p_2 + \ldots + b_s p_s \quad . \quad . \quad . \quad . \quad . \quad (3)$$

the b's being known coefficients.

The method of the solution follows at once from the statement of the problem. It will be convenient to use the notation $E(x)$ for the mean value of any variate x. Denote by θ' a linear function of x's :

$$\theta' = \Sigma(\lambda_i x_i) \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Our problem consists in determining the λ's so that both

$$E(\theta') = \theta \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and

$$\sigma^2_{\theta'} = E(\theta' - \theta)^2 = \textit{minimum} \quad . \quad . \quad . \quad . \quad (6)$$

The condition (5) leads to the following equations which we reach by taking into account (2) and (3) :

$$E(\theta') = \Sigma\lambda_i E(x_i) = \theta \quad . \quad . \quad . \quad . \quad . \quad (7)$$

or

$$\begin{aligned} &p_1\Sigma(\lambda_i a_{i1}) + p_2\Sigma(\lambda_i a_{i2}) + \ldots + p_s\Sigma(\lambda_i a_{is}) = \\ &p_1 b_1 \qquad + p_2 b_2 \qquad + \ldots + p_s b_s \quad . \quad . \quad . \quad (8) \end{aligned}$$

This equality should hold whatever be the unknown parameters $p_1, p_2, \ldots p_s$.

This is possible only if we have

$$\begin{aligned} \Sigma(\lambda_i a_{i1}) &= b_1 \\ \Sigma(\lambda_i a_{i2}) &= b_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9) \\ &\ldots \\ \Sigma(\lambda_i a_{is}) &= b_s \end{aligned}$$

As there are s linear equations with regard to $n \geqslant s$ unknown coefficients λ we may generally make a choice among all possible systems satisfying (9) in order to satisfy (6).

Using the known formula for the variance of a linear function of n variables x_i, we may write the condition (6) in the form

$$\sigma^2_{\theta'} = \Sigma(\lambda_i{}^2\sigma_i{}^2) + 2\Sigma\Sigma(\lambda_i\lambda_j\sigma_i\sigma_j r_{ij}) \quad . \quad . \quad . \quad (10)$$

The values of λ satisfying (9) and minimising (10) may be easily found provided we have sufficient information concerning the σ_i's and the correlation coefficients r_{ij}. The case considered by Markoff is when it is known that $r_{ij} = 0$, and

$$\sigma_i{}^2 = \frac{\sigma^2}{P_i} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

P_i being a known number and σ^2 some unknown constant. In this case the function (10) to minimize may be written in the form

$$\sigma^2_{\theta'} = \sigma^2\Sigma\left(\frac{\lambda_i{}^2}{P_i}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

The solution of minimizing (12), or, which is the same, of minimizing the sum

$$\sum\left(\frac{\lambda_i^2}{P_i}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

under the condition of having (9) satisfied, is now a straight forward one. It requires, however, that among n equations (2) connecting the means A_i with the parameters p there should be at least s independent equations. If this condition is satisfied, then the solution of the whole problem is given by the following theorem of Markoff.*

The best linear estimate of θ *is obtained by substituting in* (3) *instead of parameters* p, *linear functions of the* x'*s say*

$$q^\circ_1, q^\circ_2, \ldots q^\circ_s. \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

found by minimizing the sum of squares

$$S = \Sigma(x_i - a_{i1}q_1 - a_{i2}q_2 - \ldots - a_{is}q_s)^2 P_i \quad . \quad (15)$$

with regard to the q'*s considered as independent variables.*

The second part of this very important theorem gives us an estimate of the variance of θ'. Denote by μ^2 the estimate of $\sigma^2_{\theta'}$ and by S_0 the minimum value of S obtained from (15) by substituting the functions (14) for the q's. Then, according to the result of Markoff

$$\mu^2 = \frac{S_0}{n - s}\sum\left(\frac{\lambda_i^2}{P_i}\right) \quad . \quad . \quad . \quad . \quad . \quad (16)$$

If it were known that $P_1 = P_2 = \ldots = 1$, then the ratio $S_0/(n - s)$ would be the estimate of the value of the variance σ^2, common to all x's. Considering the denominator in (16) we recognize in $n - s$ the number of degrees of freedom—the concept introduced by R. A. Fisher—equal to the number of observations, n, minus the number of independent parameters.

The above results concern the case when the variables (1) are independent. If their dependency arises from the fact that they are characters of individuals randomly drawn from limited populations, then the above results may be easily adapted to this case. In particular the method of Markoff applies when we are dealing with stratified sampling.

Consider, for example, the problem of estimating the sum $\theta = \Sigma\Sigma(u)$ of some numbers u associated with the elements of some stratified population π. Denote by π_i the i-th stratum ($i = 1, 2, \ldots k$), by M_i— the number of its elements, by $\bar{u}_i$ the mean of

* Actually Markoff formulated his theorem in a slightly different form, but this difference is of no importance, and the form here chosen seems to be more convenient.

the u's corresponding to the elements of π_i and by σ_i their standard deviation. The letters u_{ij}, x_{ij}, and m_i will denote respectively the value of u corresponding to the j-th element of π_i, the value of u corresponding to the j-th of the elements selected from π_i to form the sample and the number of the elements of π_i selected for the sample. We may write now

$$\theta = \Sigma(M_i \bar{u}_i) \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

To find the best linear estimate of θ we write

$$\theta' = \sum_{i=1}^{k} \sum_{j=1}^{m_i} (\lambda_{ij} x_{ij}) \quad . \quad . \quad . \quad . \quad . \quad (18)$$

and then try to find such values of the coefficients λ which would satisfy the conditions

$$E(\theta') = \theta \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

whatever is the value of θ and of the $\bar{u}$'s and

$$E(\theta' - \theta)^2 = \textit{minimum} \quad . \quad . \quad . \quad . \quad (20)$$

Owing to (17) and to the obvious fact that

$$E(x_{ij}) = \bar{u}_i \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

the condition (19) transforms itself into the following:

$$\sum_{i=1}^{k} \left(\bar{u}_i \left(\sum_{j=1}^{m_i} \lambda_{ij} - M_i \right) \right) = 0. \quad . \quad . \quad . \quad . \quad (22)$$

This shows that the necessary and sufficient conditions which the λ's must satisfy in order to have (19) satisfied whatever the unknown properties of the population, will be the following:

$$\sum_{j=1}^{m_i} (\lambda_{ij}) = M_i \ (i = 1, 2, \ldots k) \quad . \quad . \quad . \quad (23)$$

This being fixed, consider the condition (20).

Straightforward algebra gives for the left-hand side of (20) the expression, say,

$$\sigma^2_{\theta'} = \sum_{i=1}^{k} \left\{ \sigma_i^2 \left(m_i \frac{M_i - m_i}{M_i - 1} \lambda_i^2 + \frac{M_i}{M_i - 1} \sum_{j=1}^{m_i} (\lambda_{ij} - \lambda_i)^2 \right) \right\}, \quad . \quad (24)$$

where λ_i stands for the mean value of λ_{ij} calculated from (23), that is,

$$\lambda_i = \frac{M_i}{m_i}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

We see now that $\sigma^2_{\theta'}$ is minimized by a system of λ's satisfying (23) in which, whatever $j = 1, 2, \ldots m_i$

$$\lambda_{ij} = \lambda_i = \frac{M_i}{m_i}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (26)$$

It will be noticed that this result holds good not only whatever be the unknown $\bar{u}_i$, but also whatever the standard deviations σ_i. Thus if we denote by $\bar{x}_i$ the mean value x_{ij} for any stratum, then the function

$$\theta' = \sum_{i=1}^{k} (M_i \bar{x}_i) \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

is the best linear estimate of θ *whatever the properties of the population.* The familiar formula for the variance of θ' is easily obtained from (24) and (25):

$$\sigma^2_{\theta'} = \sum_{i=1}^{k} M_i \frac{M_i - m_i}{m_i} \frac{M_i \sigma_i^2}{M_i - 1} \quad . \quad . \quad . \quad (28)$$

The estimate θ' is, of course, the one which could be suggested on purely intuitive grounds. The advantage of using Markoff's method consists (i) in avoiding biased estimates, which may be sometimes used when their choice is based on intuition only, and (ii) in finding *the best* linear estimates. It would be rather difficult to fulfil this last condition on intuitive grounds only.

Note III. The consistency and the efficiency of the estimate of C. Gini and L. Galvani.

Consider a population π of districts divided into second order strata, π_{yv}, according to the values of the control y and the number v of individuals in the districts. Thus any district in the stratum π_{yv} contains the same number of individuals v and the value of the control corresponding to each district is also the same y. Denote by M_{yv} and m_{yv} the total number of districts contained in π_{yv} and the number of them to be included in the sample. The letters u_{yvi} and x_{yvi} will denote the values of the character sought, x, associated respectively with the i-th district of the stratum π_{yv} and with the i-th district out of the m_{yv} of them, which have been selected from this stratum. The letters u_{yv} and x_{yv} will denote the means of u_{yvi} and x_{yvi} corresponding to the stratum π_{yv} and to the partial sample of districts, drawn from this stratum. The standard deviation of u_{yvi} will be denoted by σ_{yv}. Finally, X and Y will denote the weighted means of the character sought x and of the control, calculated for the whole population π. The sample weighted means will be denoted by X_Σ and Y_Σ. Denote further by

$$W = \sum_y \sum_v (M_{yv} v) \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

We shall have

$$X = \frac{1}{W}\sum_{y}\sum_{v}(M_{yv}vu_{yv}), \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$Y = \frac{1}{W}\sum_{y}\sum_{v}(M_{yv}vy), \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

$$X_{\Sigma} = \frac{\sum_{y}\sum_{v}(m_{yv}vx_{yv})}{\sum_{y}\sum_{v}(m_{yv}v)}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

$$Y_{\Sigma} = \frac{\sum_{y}\sum_{v}(m_{yv}vy)}{\sum_{y}\sum_{v}(m_{yv}v)}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Using this notation, we shall now consider the necessary condition which must be satisfied by x and y in the case when the estimate of Gini and Galvani is consistent. This condition is more easily found when we consider the ideal case where the sample weighted mean Y_{Σ} is exactly equal to its population value Y. In this case the estimate of Gini and Galvani reduces itself to X_{Σ}. Thus we shall consider the conditions under which the mean of X_{Σ} in repeated samples is equal to X whenever

$$Y_{\Sigma} = Y. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

We suppose that the numbers m_{yv} are fixed in some way or other in order to satisfy (6), and consider the mean value of X_{Σ}, which we should get from all possible samples, corresponding to the fixed values of m_{yv}. We shall have

$$E(X_{\Sigma}) = \frac{\sum_{y}\sum_{v}(m_{yv}vu_{yv})}{\sum_{y}\sum_{v}(m_{yv}v)} = \overline{X}_{\Sigma} \text{ (say)}. \quad . \quad . \quad . \quad (7)$$

Now we wish to have

$$\overline{X}_{\Sigma} = X, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

whenever (6) is satisfied.

The equations (6) and (8) may be written in the following form

$$\sum_{y}\sum_{v}m_{yv}v(y - Y) = 0, \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$\sum_{y}\sum_{v}m_{yv}v(u_{yv} - X) = 0, \quad . \quad . \quad . \quad . \quad . \quad (10)$$

and it is easily seen that if it is required that (10) holds good whenever the numbers m_{yv} satisfy (9), it is necessary that

$$u_{yv} - X = A(y - Y), \quad . \quad . \quad . \quad . \quad . \quad (11)$$

A being an absolute constant, independent of y and v.

The necessary (and obviously also sufficient) condition (11) of the consistency of the estimate of Gini and Galvani is a rather peculiar one. It is to be noticed that it is more limiting than the

hypothesis H mentioned in the text (p. 571). In fact (11) means that the regression of x on y should be linear not only if we consider the whole population of districts, but also if we consider the part of this population containing only districts with a fixed number of individuals v. Furthermore, the regression lines corresponding to different v's should be the same. Denote the hypothesis that these conditions are fulfilled by H'. Obviously if H' is true, then H is also true, but not inversely.

It is perhaps worth noticing that X_Σ may be a consistent estimate of X for a special system of the m's. In fact denote

$$\sum_y \sum_v (m_{yv} v) = w \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

and subtract (2) from (7):

$$\overline{X}_\Sigma - X = \sum_y \sum_v \left\{ v u_{yv} \left(\frac{m_{yv}}{w} - \frac{M_{yv}}{W} \right) \right\} \quad . \quad . \quad . \quad (13)$$

It is easily seen that if for any y and v

$$\frac{m_{yv}}{M_{yv}} = \frac{w}{W} = \text{const.}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

then (13) vanishes and X_Σ becomes an unbiased estimate of X whatever the properties of the population. However, it may be noticed that the fulfilment of (14) means the rejection of the principle of purposive selection.

Assume now that the hypothesis H' is true and find the best linear estimate of X corresponding to any system of m's not necessarily satisfying the condition (6). As the values x_{yvi}, x_{yvj} corresponding to two districts drawn from the same stratum π_{yv} are correlated, it is impossible to apply the theorem of Markoff at once. We shall do so later on. However, we must start by following the general method. Denoting by θ' the linear function of x's required, we shall have

$$\theta' = \sum_y \sum_v \sum_i (\lambda_{yvi} x_{yvi}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

It must satisfy the conditions

$$E(\theta') = X \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

and if $\sigma^2_{\theta'}$ stands for the variance of θ',

$$\sigma^2_{\theta'} = \textit{minimum}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

Since the values of the control are known for all districts, we may change the origin of co-ordinates and assume that $Y = 0$.

According to the assumed hypothesis, H', (see (11)) we have then

$$E(x_{yvi}) = u_{yv} = X + Ay \quad . \quad . \quad . \quad . \quad . \quad (18)$$

where X and A play the rôle of the unknown parameters, p, considered in the theory of the Markoff method. Therefore

$$E(\theta') = \sum_{y}\sum_{v}\sum_{i}(\lambda_{yvi}(X + Ay)) = X\sum_{y}\sum_{v}\sum_{i}(\lambda_{yvi}) + A\sum_{y}\{y\sum_{v}\sum_{i}(\lambda_{yvi})\} \quad . \quad . \quad . \quad (19)$$

The values of the λ's must be so chosen that this last expression should be identically equal to X whatever the unknown value of A. Thus we should have, say

$$f = \sum_{y}\sum_{v}\sum_{i}(\lambda_{yvi}) = 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

$$\phi = \sum_{y}(y\sum_{v}\sum_{i}(\lambda_{yvi})) = 0 \quad . \quad . \quad . \quad . \quad . \quad (21)$$

These are the conditions of the consistency of the estimate θ'. Now consider the condition of its being a best linear estimate. Taking into account the correlation between x_{yvi} and x_{yvj}, the variance of θ' may be written in the form, which is easy to check

$$\sigma^2_{\theta'} = \sum_{y}\sum_{v}\left\{\sigma^2_{yv}\left(\sum_{i}(\lambda^2_{yvi}) - \frac{2}{M_{yv} - 1}\sum\sum(\lambda_{yvi}\lambda_{yvj})\right)\right\} \quad . \quad (22)$$

We proceed to minimize this expression with regard to all systems of the λ's satisfying (20) and (21). For this purpose we shall equate to zero the derivatives of the function, say,

$$F = \sigma^2_{\theta'} - 2\alpha f - 2\beta\phi, \quad . \quad . \quad . \quad . \quad . \quad (23)$$

α and β being some coefficients to be determined from (20) and (21). We have

$$\frac{\partial F}{\partial \lambda_{yvi}} = 2\sigma^2_{yv}\left\{\lambda_{yvi} - \frac{1}{M_{yv} - 1}\left(\sum_{j=1}^{m_{yv}}(\lambda_{yvj}) - \lambda_{yvi}\right) - \alpha - \beta y\right\} = 0 \quad (24)$$

The above equation may be written in the following form

$$M_{yv}\lambda_{yvi} = \sum_{j=1}^{m_{vy}}(\lambda_{yvi}) + (\alpha + \beta y)(M_{yv} - 1), \quad . \quad . \quad (25)$$

which shows that whatever $i = 1, 2, \ldots m_{yv}$, the control y and v being fixed, λ_{yvi} has a constant value, say λ_{yv}. This value may be obtained from (25), in terms of α and β

$$\lambda_{yvi} = \lambda_{yv} = (\alpha + \beta y)\frac{M_{yv} - 1}{M_{yv} - m_{yv}} \quad . \quad . \quad . \quad (26)$$

Substituting this value into (20) and (21) we should be able to calculate α and β and then to find λ_{yv} from (26). However, it is easier to get the result by using the Markoff theorem.

The results which are already obtained justify us in writing

$$\theta' = \sum_y \sum_v (\lambda_{yv} \sum_i (x_{yvi})) = \sum_y \sum_v (\lambda_{yv} m_{yv} x_{yv}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

and thus to treat θ' as a linear function of the variables x_{yv} which, being associated with samplings from different strata, are totally independent. The mean value of x_{yv} is the same as of x_{yvi}, equal to u_{yv}. The variance of x_{yv}, say S^2_{yv}, is connected with the variance of x_{yvi} and is as follows :

$$S^2_{yv} = \frac{M_{yv} - m_{yv}}{m_{yv}(M_{yv} - 1)} \sigma^2_{yv} = \frac{1}{Q_{yv}} \text{ (say)}. \quad . \quad . \quad (28)$$

Thus the original problem of finding a linear function θ' of *dependent* variables x_{yvi} satisfying (20) and (21) is now reduced to that of finding a linear function (27) of *independent* variates x_{yv} satisfying the same conditions. This may be done by applying the Markoff theorem. However, we shall have to assume some additional hypotheses concerning the variances σ^2_{yv}. Gini and Galvani assumed in their paper a hypothesis (p. 63) which, I think, in my terminology and notation could be expressed by the equality

$$\sigma^2_{yv} = \sigma^2 = \textit{constant}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (29)$$

There is no difficulty in assuming this hypothesis and in finding the best linear estimate of X directly from the Markoff theorem. Comparing (28) and (29) we see that the rôle of the " weights," P, involved in the Markoff theorem is now played by the ratios, say,

$$P_{yv} = \frac{m_{yv}(M_{yv} - 1)}{M_{yv} - m_{yv}} \quad . \quad . \quad . \quad . \quad (30)$$

Thus the best linear estimate of X is found by minimizing the sum of squares

$$S = \sum_y \sum_v \left\{ (x_{yv} - q_1 - q_2 y)^2 \frac{m_{yv}(M_{yv} - 1)}{M_{yv} - m_{yv}} \right\} \quad . \quad . \quad (31)$$

with regard to q_1 and q_2 considered as independent variables. The best linear estimate sought, θ', will be the solution for q_1, minimizing (31). It is easily seen that it is not the one which has been suggested by Gini and Galvani, which results from minimizing the sum of squares, say

$$S' = \sum_y \sum_v \{(x_{yv} - q_1 - q_2 y)^2 v m_{yv} \quad . \quad . \quad . \quad (32)$$

The estimate of Gini and Galvani becomes the best linear estimate when we assume the hypothesis, which will be denoted by H_1, that the variances S^2_{yv} are inversely proportional to vm_{yv}. This could be

true, for instance, if the districts contained in each stratum were samples of v individuals drawn randomly from the population II. In this case we should have

$$\sigma^2_{yv} = \frac{\sigma^2_y}{v} \text{ (say).} \quad . \quad . \quad . \quad . \quad . \quad . \quad (33)$$

Additional assumptions, that σ^2_y is independent of y, and is equal to σ^2, say, and that each $m_{yv} = 1$ will reduce the formula (28) to

$$S^2_{yv} = \frac{\sigma^2}{v}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (34)$$

which will lead to minimizing (32) when getting the best linear estimate of X. I think it would be exceedingly difficult to find instances in social, vital or economic statistics in which the hypothesis H_1 would be true. However, it may be true in some engineering problems.

I want to emphasize that the general problem of estimation of any given collective character θ of a population π must be considered from two different points of view: (i) Given a sample from the population, obtained in some known manner, what arithmetical procedure will give us an unbiased and a most accurate estimate of θ? (ii) What method of sampling will give samples, allowing the most accurate estimates? These two aspects of the problem may be traced in any theoretical research concerning the representative method. Often the solutions proposed are based only on intuition and require theoretical justification. The principle of the purposive selection method, advising selection of samples such that the weighted sample means of the controls should be equal to their population values, is an intuitive solution of the problem (ii). The methods which have been proposed to estimate the unknown weighted population mean X are the solutions of the problem of kind (i). In the first part of the present Note I have considered the conditions under which the intuitive solutions of the problem (i) are justified. Now I shall proceed to consider whether, and if so then under what conditions, is justified the solution of the problem of kind (ii).

To do so I shall assume that the hypothesis H' is satisfied and shall consider the variance of the best linear estimate, θ', of the unknown weighted mean X. Its expression will involve the numbers m_{yv} of districts selected for the sample from each stratum π_{yv}. It will be possible to see what system of these numbers m_{yv} would minimize the value of the variance $\sigma^2_{\theta'}$ of θ'. We shall see that under certain, rather limiting, conditions, the system of m_{yv} minimizing $\sigma^2_{\theta'}$ will be the system for which the sample weighted mean of the control is equal to its population value.

We shall consider the question in its full generality and shall make no assumptions about the variances, σ^2_{yv}, of the character sought within the second order strata. This will lead to the expression (28) for the variances of numbers x_{yv}. The best linear estimate θ' of X will be obtained by minimizing the sum of squares

$$S = \sum_y \sum_v \{(x_{yv} - q_1 - q_2 y)^2 Q_{yv}\} \quad . \quad . \quad . \quad (35)$$

with regard to q_1 and q_2. The solution is easily obtained and is given by the formula

$$\theta' = \sum_y \sum_v (\lambda_{yv} x_{yv}), \quad . \quad . \quad . \quad . \quad . \quad . \quad (36)$$

where

$$\lambda_{yv} = \frac{\Sigma\Sigma(Q_{yv}y^2) - y\Sigma\Sigma(Q_{yv}y)}{\Sigma\Sigma(Q_{yv}y^2)\Sigma\Sigma(Q_{yv}) - (\Sigma\Sigma(Q_{yv}y))^2} Q_{yv} \quad . \quad . \quad (37)$$

The variance of (36) is given by the familiar formula

$$\sigma^2_{\theta'} = \Sigma\Sigma(\lambda^2_{yv} S^2_{yv}). \quad . \quad . \quad . \quad . \quad . \quad . \quad (38)$$

This, owing to (37) and (28) reduces itself to the following

$$\sigma^2_{\theta'} = \frac{1}{\Sigma\Sigma(Q_{yv})}\left\{1 + \frac{(\Sigma\Sigma(Q_{yv}y))^2}{\Sigma\Sigma(Q_{yv}y^2)\Sigma\Sigma(Q_{yv}) - (\Sigma\Sigma(Q_{yv}y))^2}\right\} \quad . \quad (39)$$

Considering this formula, we see that it will provide a small value of $\sigma^2_{\theta'}$ if we succeed in minimizing say

$$| Y'_\Sigma | = \frac{| \Sigma\Sigma(Q_{yv}y) |}{\Sigma\Sigma(Q_{yv})} \quad . \quad . \quad . \quad . \quad . \quad (40)$$

and at the same time maximize $\Sigma\Sigma(Q_{yv})$. The former expression (40) may be considered as a weighted mean of the control, calculated for the sample, where the Q_{yv} are playing the rôle of the weights. Remembering that by a proper choice of the origin of co-ordinates, we have reduced the value of the population weighted mean

$$Y = \frac{\Sigma\Sigma(M_{yv}vy)}{\Sigma\Sigma(M_{yv}v)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (41)$$

to zero, we see that the above result would justify the principle of purposive selection, if the new weights Q_{yv} were proportional to the weights used in (4); thus if we had

$$Q_{yv} = \frac{m_{yv}(M_{yv} - 1)}{(M_{yv} - m_{yv})\sigma^2_{yv}} = \frac{m_{yv}v}{C} \quad . \quad . \quad . \quad (42)$$

C being a constant. Solving (42) with regard to C we get

$$C = \frac{M_{yv} - m_{yv}}{M_{yv} - 1} v\sigma^2_{yv} \quad . \quad . \quad . \quad . \quad . \quad (43)$$

It is easily seen that the right-hand side of (43) may be constant

when $m_{yv} = 1$ and if the values of σ^2_{yv} are independent of y and inversely proportioned to v; thus when the hypothesis H_1 is satisfied. Obviously this is not the only condition under which (43) is constant, but it is difficult to formulate any other hypothesis which would concern the unknown values of σ^2_{yv}.

The analysis of the formula (39) might be carried a little further, but the considerable size of the present Note and the results already obtained suggest that this may be superfluous.

The conclusions which were obtained above may be summed up as follows:

(*a*) The estimate of Gini and Galvani is unbiased only if the very limiting hypothesis H' is satisfied.

(*b*) This estimate is the best linear estimate when another still more limiting hypothesis H_1 is satisfied. This hypothesis consists in the assumption that the variation of x between districts included in each second order stratum, π_{yv}, depends only upon the value of v and is such as could arise if the districts were random samples from the population studied. This condition is hardly ever satisfied.

(*c*) If the hypothesis H' is true, then the principle of purposive selection of districts so as to keep the value of the sample weighted mean of the control equal to its population value, is justified when the hypothesis H_1 is true. The dependence of the weighted mean on the system of weights being only slight, it may be assumed that this principle is approximately satisfied even if the hypothesis H_1 is not exact. Other conclusions which may be considered as corrections to the principle of purposive selection, may be drawn from the formula (39).

Whether it is likely that the hypotheses H' and H_1 are justified and whether the divergencies in this respect will seriously influence the accuracy of the results of the application of the purposive selection method, must be considered in any special case.

It will be useful to finish this Note by a numerical illustration of the assumptions involved in the hypotheses H, H', and H_1. The following three tables give data concerning three populations π_1, π_2, and π_3 of districts. Each population is subdivided into two first order strata, according to the values of the control $y = -1$ and $y = +1$. Each first order stratum is in its turn subdivided into two second order strata according to the size of the districts $v = 1$ and $v = 6$. (Obviously the units in which the number v of persons included in a district is measured, is of no importance. One unit may be, for instance, 10,000 individuals.) The Tables I, II, III, give the values of the character x for each district in a stratum (these, according to notation used in this Note, are denoted by u_i), their arithmetic mean $\bar{u}$, and their variance σ^2.

TABLE I.

Population π_1.

$y = -1$.		$y = +1$.	
$v = 1$. Stratum I.	$v = 6$. Stratum II.	$v = 1$. Stratum III.	$v = 6$. Stratum IV.
$u_1 = -4$ $u_2 = -1$ $u_3 = +2$	$u_1 = -2$ $u_2 = 0$ —	$u_1 = -2$ $u_2 = +1$ $u_3 = +4$	$u_1 = 0$ $u_2 = 2$ —
$\bar{u} = -1$ $\sigma^2 = 6$	$\bar{u} = -1$ $\sigma^2 = 1$	$\bar{u} = +1$ $\sigma^2 = 6$	$\bar{u} = +1$ $\sigma^2 = 1$

TABLE II.

Population π_2.

$y = -1$.		$y = +1$.	
$v = 1$. Stratum I.	$v = 6$. Stratum II.	$v = 1$. Stratum III.	$v = 6$. Stratum IV.
$u_1 = -2$ $u_2 = 0$ —	$u_1 = -4$ $u_2 = -1$ $u_3 = +2$	$u_1 = 0$ $u_2 = +2$ —	$u_1 = -2$ $u_2 = +1$ $u_3 = +4$
$\bar{u} = -1$ $\sigma^2 = Q^{-1} = 1$	$\bar{u} = -1$ $\sigma^2 = Q^{-1} = 6$	$\bar{u} = +1$ $\sigma^2 = Q^{-1} = 1$	$\bar{u} = +1$ $\sigma^2 = Q^{-1} = 6$

TABLE III.

Population π_3.

$y = -1$.		$y = +1$.	
$v = 1$. Stratum I.	$v = 6$. Stratum II.	$v = 1$. Stratum III.	$v = 6$. Stratum IV.
$u_1 = u_2 = u_3 = -2$ $u_4 = u_5 = u_6 = 0$ —	$u_1 = -6$ $u_2 = -3$ $u_3 = 0$	$u_1 = u_2 = u_3 = 0$ $u_4 = u_5 = u_6 = 2$ —	$u_1 = 0$ $u_2 = 3$ $u_3 = 6$
$\bar{u} = -1$ $\sigma^2 = 1$	$\bar{u} = -3$ $\sigma^2 = 6$	$\bar{u} = +1$ $\sigma^2 = 1$	$\bar{u} = 3$ $\sigma^2 = 6$

It will be seen that for all three populations $X = Y = 0$. Owing to the fact that y has only two different values, the regression of x on y is linear in all populations, and thus all of them satisfy the hypothesis H. The populations π_1 and π_2 satisfy also the hypothesis H'. In fact, the means $\bar{u}$ in the second order strata do not depend upon the value of v. It is not so in the population π_3. Here in strata corresponding to $v = 1$ the regression coefficient of x on y is

equal to $A_1 = 1$ and in strata corresponding to $v = 6$, to $A_2 = 3$ Thus the population π_3 does not satisfy the hypothesis H'.

The hypothesis H_1 is satisfied only in the population π_1 as here the variances σ^2 are inversely proportional to the v's.

Consequently, whatever be the numbers, say m_1, m_2, m_3, m_4 of districts selected for the sample from the four strata of the populations π_1 and π_2 so that the sample weighted mean of the control y is equal to zero, the estimate of Gini and Galvani will have its mean in repeated samples equal to $X = 0$. If it were $m_1 = m_2 = m_3 = m_4 = 1$, then in the case of the population π_1 the estimate of Gini and Galvani would be the best linear estimate. Its variance is equal to $\frac{3}{7}$. When sampling in the same way from the population π_2 the estimate of Gini and Galvani would not be the best linear estimate. In fact its variance would be equal to $\frac{31}{14}$. On the other hand, the best linear estimate, which could be derived from the formulæ (36) and (37), namely,

$$\theta' = \frac{\Sigma(Qx)}{\Sigma(Q)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (44)$$

would have the variance $\frac{3}{7}$, as previously.

As the hypothesis H' is not satisfied in the population π_3, it is possible to find such a system of the m's that the estimate of Gini and Galvani will have its mean value in repeated samples not equal to $X = 0$. Owing to the exceptional symmetry of the population π_3, there will be many systems of m's by which the estimate of Gini and Galvani will be consistent. However, let us consider the system $m_1 = 6$, $m_2 = 0$, $m_3 = 0$, $m_4 = 1$. Obviously the sample weighted mean of the control $Y_2 = 0$. There will be only three possible samples corresponding to the fixed system of the m's depending upon the choice of the district in the fourth stratum. The mean of the estimates of Gini and Galvani, calculated for these three samples will be

$$\bar{X} = \frac{(6)(-1) + (6)(3)}{12} = 1 \quad . \quad . \quad . \quad . \quad (45)$$

and is not equal to $X = 0$. Thus for this system of the m's the estimate of Gini and Galvani is not consistent. It is easily seen that if there were more districts in each stratum the number of systems, for which the estimate of Gini and Galvani would not be consistent, would be increased. It would be also considerably larger if the structure of the population π_3 were not symmetrical.

Discussion on Dr. Neyman's Paper

Professor Bowley: There are some who appear to pride themselves on their absence of knowledge of mathematics. I never understood why it should be a matter of pride. I do not think, however, that there are very many who now hold that mathematics is not properly appropriate to the study of statistical problems. This paper will, when it is thoroughly studied, do very much to remove any remaining doubt that the mathematical approach is of fundamental importance. Sampling is at the very root of a great deal of statistical investigation and, as Dr. Neyman points out, of increasing use and applicability; it is therefore of the first practical importance to decide what is the best method of sampling—best in the sense that the best use will be made of the resources at the disposal of the investigator, and that there shall be—if the two are consistent—a minimum expenditure of time.

This paper of Dr. Neyman's will be found to answer most of the questions which relate to the setting out of an investigation by sample. One of the things that is so interesting in it is the analysis of the problems. There is not one perfect method of sampling; the method depends upon the nature of the material which is available or which can be obtained. To me a new suggestion in the classification is this stratified random sampling of groups. In the analysis to which he has referred in much too favourable terms, I had distinguished in the ways he named certain methods of sampling, but in effect I realize that it is precisely the method which he has discussed that I have been driven by circumstances to use or to recommend.

In the Survey of London, the unit was not the family or the person, but the house—the unit which was provided by the directories we had. In the recommendation I had the honour to make with Mr. Robertson to the Government of India, of sampling on a very large scale, the unit suggested is the village. In the recommendations, which had no effect, which I made with regard to the 1931 Population Census, of a method which I had in fact made use of on the 1911 figures, the unit was the householder's schedule, and I suggested that if one took one household in five hundred throughout the country we could deduce some of the results of the Census which now, after three years, we are still waiting for. But that, I think, is not necessarily the method appropriate to all problems, nor do I understand Dr. Neyman to recommend it universally.

I am surprised that he thought that when in 1925 I examined the problem of representative sampling for the International Institute of Statistics, I gave equal importance to that method, as I defined it, and to others. Certainly I thought I damned it with very faint praise at the end of the summary of my report. I agree that it is difficult to formulate, difficult to carry out, and I still think that it is very difficult to get a good estimate of the precision of the result, except in rather unusual cases.

The second problem, after discussing the material and defining the best method of sampling, is to get a definite estimate of the precision of the sampling in the sense that when one has a result, one

knows that it can be trusted, without defining that term, to 1 per cent. or 1 shilling, or whatever it may be ; and it is partly because of that necessity that a common method, partly intuitive, of choosing the sample from the obvious—of taking the mode rather than anything else—is hopelessly faulty, because not only is there difficulty in obtaining an average, but there is no means that I know of obtaining precision.

This method of stratification, so far as it differs in precision from purely random sampling, gives an improvement in precision. If, in fact, one has made a stratified selection and writes down the statement of the precision as if it had been purely random, then one is on the safe side. I have myself generally been content to let it go at that for two reasons : (1) that it is very difficult to measure the additional precision due to stratification in ordinary material—at any rate it is a very lengthy business ; and (2) (of quite a different nature) I have had to explain and try to justify the methods of sampling to non-technical readers, and therefore I have been obliged to leave out a great deal that Dr. Neyman would have put in. But I have endeavoured to be on the safe side, and what I have neglected I know to be unimportant.

This process of stratification can be thought of or applied in various ways. In London we took the simplest way, taking one house in thirty, forty, or fifty, right through the directory, so that the houses we examined, if plotted on a map of London, would be regularly distributed in proportion to the density of the area and every region would be included. But in that way, I think it must appeal to everyone who studies the problem.

A new point that comes out in the paper is that a selection of the units to be examined for one purpose is not necessarily the best for other purposes. As the selection has generally to be made once and for all, that becomes an important consideration in selecting the method to pursue. I am glad that the general recommendation for this kind of purpose is this random stratified sampling.

To give an example in this case in London, if we take a house, we take one, two, or three families. The family is not selected at random, and within the family the persons are not selected at random. If co-existence in one case gives positive correlation in one instance, it may be negative in another, and similarly with relationship between persons in the family. The effect of this is difficult to estimate. My point, however, would be that we must secure a selection that would minimize the influence of these unknown correlations, and one method would be to make our samples sufficiently large to make them unimpaired by the most unfavourable hypothesis.

After Dr. Neyman's very courteous references to my work on the subject, it is somewhat ungrateful that I feel it my duty to criticize the theory of probabilities in Section II, part 1, and I am very glad Professor Fisher is present, as it is his work that Dr. Neyman has accepted and incorporated. I am not certain whether to ask for an explanation or to cast a doubt. It is suggested in the paper that the work is difficult to follow and I may be one of those who

have been misled by it. I can only say I have read it at the time it appeared and since, and I read Dr. Neyman's elucidation of it yesterday with great care. I am referring to Dr. Neyman's confidence limits. I am not at all sure that the " confidence " is not a " confidence trick." Put in a simple form I think the method is as follows :—Given that in a sample of 1,000 taken at random, there are 1 in 10 with the defined quality, and given that the population from which the sample was drawn contained any proportion between 120 and 80 per thousand, then the chance of such an occurrence is less than one in twenty (approx.). Actual figures, of course, do not matter. That margin between 120 and 80 per thousand in the assumed population is shown on the vertical of the confidence belt in the very illuminating graphs which Dr. Neyman has given. Does that really take us any further? Do we know more than was known to Todhunter? Does it take us beyond Karl Pearson and Edgeworth? Does it really lead us towards what we need—the chance that in the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event has occurred *or* the proportion in the population is within the limits. To balance these things we must make an estimate and form a judgment as to the likelihood of the proportion in the universe—the very thing that is supposed to be eliminated. I do not say that we are making crude judgments that everything is equal throughout the possible range, but I think we are making some assumption or we have not got any further. I do not know that I have expressed my thoughts quite accurately, but it is not a thing that has occurred to me for the first time this evening ; it is the difficulty I have felt since the method was first propounded. The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity.

I regret that in opening up that subject I have distracted attention from Dr. Neyman's paper, but since he has made that an integral part of his paper, I think it a proper occasion on which to make this kind of statement.

With reference to my formula, quoted in Section IV, equation 24, I must admit that the original passage is obscure. In Dr. Neyman's notation and with only one control, my estimate would be *

$$X^1 = X^2 + (r_{ax} - r_{ay}r_{yx})\sigma_x \,.\, \sigma_a/\bar{a}.$$

where the a's are the weights attached to the x's in the weighted average. The second term is zero, if there is no correlation between the weights and the divergencies, e's, from the linear regression equation.† In other cases, K should be regarded as attached to the error term negatively, since the weighted average of the e is not zero, but $-K$. The formula is then " consistent." I am not, however,

* The complete formula in the notation I used may be written

$$X = X_u + \{\sigma_x \,.\, R_x/R_{11} \,.\, \sigma_a/\bar{a} - E\}$$

where
$$R_x = \begin{vmatrix} r_{ax}r_{au}r_{av} \ldots r_{az} & \rho_{av} \ldots r_{vz}\rho_{av} \\ \vdots\ \vdots\ \vdots \ldots & \vdots \ldots \end{vmatrix}$$

† There is considerable correlation in Table I in Dr. Neyman's paper.

at all sure that the particular hypotheses underlying my treatment are the best; it was in some way pioneer work, and I should have been astonished if no improvement should have been made in course of time.

I wish to propose a hearty vote of thanks to Dr. Neyman for his very important paper.

Dr. E. S. Pearson: I have great pleasure in seconding this vote of thanks to Dr. Neyman and welcoming him here among us to-day. I think we are all very glad that as one of our Fellows from abroad he has been able to take the opportunity of being in England to read a paper before us.

I should like to try to express in a few words what appears to me to be the essential contribution to statistical method that Dr. Neyman has made in this paper and in other work that he has done. In the past thirty or forty years the development of mathematical statistics has been an extremely rapid one; it has been associated with all the excitement of discovery, the discovery of the power of new tools in the solution of a great variety of problems. But in this rapid progress intuition was sometimes at fault; the tools used were not always the best tools, nor was it always very clear what was the meaning of these tools, nor why one tool should be used rather than another. In the last few years there has been a determined effort to clear away some of this uncertainty from our statistical reasoning. The process is not complete; it is still to some extent in the stage of controversy and discussion, but there are fundamentals that are emerging surely and steadily.

In this process many of us owe a great deal to Professor R. A. Fisher for the stimulus we have gained from wrestling with the ideas he has put forward. If I purposely use the word "wrestling," Professor Fisher will, I think, take no exception when I add that the stimulus is all the greater because it has been necessary to wrestle.

Stimulated by these ideas, as he has frankly admitted, Dr. Neyman has brought a very real contribution of his own into the field of statistical inference. For example, although in the present paper it may be regarded as only a side issue of the main subject, the approach to the problem of estimation outlined on pp. 563–567 and in Appendix I is something of very great interest. This conception of the problem of estimation is not exactly Professor Fisher's conception, but it seems to me that some of the interest lies in just those points where there are differences. I do not, however, think that this is the right place to discuss the doubts regarding confidence or fiducial intervals raised by Professor Bowley; they need to be cleared up, but that perhaps can best be done with pencil and paper at a table.

Returning to the main subject of the paper, I think the chief emphasis lies on the importance of logical planning in any investigation; the particular problem considered is that of estimating certain characteristics of a heterogeneous population from a limited sample. In doing that we have to consider, if we can, how best to take a sample, how to obtain our estimate, and what measure of

reliability to place on that estimate. To answer these questions certain assumptions regarding the unknown population are necessary, and it is important to employ a method of sampling and of estimation which will reduce these unavoidable assumptions to a minimum, and at the same time to make perfectly clear their precise implication. I do not think that this problem can be solved in any way except by the introduction of mathematics. Of course mathematics alone are not adequate, but I believe the highest level of statistical craftsmanship is only reached when, in planning an investigation, we attempt to formulate in precise, and therefore mathematical, terms the framework of hypothesis on which our final inference is to be based. It is towards this level that Dr. Neyman is pointing the way.

The special problem of the paper and other investigations of this kind has been to estimate the average value of some character in the population. This is the form of sampling problem that has presented itself to many investigators.

I would like, however, before I sit down to suggest that there are many cases in which a method of representative sampling is needed, where the average is not going to be sufficient; where it is necessary, in fact, to estimate in some way the nature and degree of heterogeneity in the population. This is perhaps not an easy problem, as we have first to determine what is the most appropriate measure of heterogeneity in a particular case.

Let me illustrate this by shifting from human populations to a population of bricks in a kiln. Owing to the arrangement of firing, the quality of the bricks varies very considerably from one part of the kiln to another. We measure various characters of those bricks, one being their strength; this though of less importance in itself is correlated with important properties of weathering. What do we want to know about the batch of bricks from the kiln? The precise average strength is of much less importance than the uniformity and in particular the lowest strengths which may be met. In determining what procedure of sampling to employ, we have first to decide what index or indices of uniformity will be of most value to us, and then to decide by what rules of sampling and calculation the most reliable estimates of these indices can be obtained.

Another illustration is that of fertilizers which may be sold in bags, the bags being drawn from a large silo in which the material is stored. Since a time element is present in the filling of the silo, there may be lack of uniformity in the quality of material. Here I think the percentage constituents of the standard fertilizers are laid down by law; for example, it may be that the nitrogen content must not be less than, say, 12 per cent. The producer wants, therefore, to be sure that he will not be summoned because the quality in one bag is found to be below that level. It is important for him to get a measure of the average nitrogen content, but it is also necessary for him to have some sampling scheme which will give him reasonable assurance that he is not sending out bags in which the content is less than a certain amount, and also that there are none in which it is too high. He must find some representative method of testing.

A final case is that of cement. There the ordinary test is to take portions of the cement from different parts of the bulked mass, mix and quarter them, and finally submit a small sample to chemical analysis. The result will give an idea of average quality but nothing more. Yet what is of most importance to the user, because it affects his technique in the mixing and setting of concrete, is the uniformity of the material. Again, therefore, some method of representative sampling is wanted which will give some idea of uniformity, not only of average.

These are problems of the future, but I believe of not so very far distant a future. In solving them a method of representative sampling should be planned, based on a sound statistical framework. This does not mean, of course, that the mathematics will be presented to the manufacturer or British Standards Institution, any more than they were presented by Dr. Neyman to the Polish Institute for Social Problems, but I hope they will be there underneath as a foundation.

Dr. Isserlis said that it was related of Sylvester that while he was Savilian Professor of Geometry at the University of Oxford, when he had discovered some recondite result, he would walk out of his rooms, dressed adequately or otherwise, buttonhole the first milkman or postman he met in the street, and hold on to that button until his victim confessed that he had been convinced by Sylvester of the truth of the theory he had discovered. Dr. Neyman had combined in one paper views on the philosophical foundations of statistical method with an exposition of an important technical problem. Each of them would have sufficed for a paper, but it was his choice to treat the two in one.

Professor Bowley, who was an expert in both, and who had been a teacher of a generation both in the theory and application of statistics, and who had been particularly the exponent of this technical problem, had referred only very briefly to the second subject, which happened to come first in the paper.

Dr. Isserlis felt that it was necessary for someone to face up to the task of trying, without the use of any mathematics, to say what it was that was found to be puzzling in this kind of philosophy, and to see where they stood. The Society had had experience in that very room of people who were more fortunate than the mathematicians. The economists, men like Mr. Hawtrey, who were accustomed in their scientific life to do without the shorthand of a special technique such as mathematics, came and spoke to their colleagues, who were experts in finance and economics, without the use of technical language, and some of their friends sat aghast at such mastery, whereas their colleagues like Sir Alfred Flux and Mr. Macrosty lapped it up like milk.

He would like briefly to refer to the thing that was worrying him, and he would try to make it clear without mathematics. There was a classical theory of probability, based on certain definitions and experiences, which told how to measure and express one's lack of full conviction in certain matters. There were two things: one might

know what was the state of affairs in a general population, and ask what was one likely to get in particular cases? One might be convinced, for instance, that pennies such as were provided by the Mint were fairly symmetrical, and on the basis of that it might be said that the theoretical probability of heads was so and so, say 50 per cent. It might then be said, supposing a coin was tossed 100 times, what should one expect? Should there be surprise if there were only 30 heads, or if there were 90 heads, or should one expect to get about 50? Without any mathematical technique it was perhaps sufficient there to say that there was an *a priori* probability of one-half.

But there was the converse problem. A coin had been tossed 100 times and fallen heads 100 times. What kind of a universe of coins had it come from? Had it come from that kind of universe of coin in which the side with the head was more likely to show up than the side with the tail?

The classical theory provided us with a method which, when limited in its application to the field for which it was intended, was perfectly legitimate. According to this theory, if we knew the *a priori* probabilities, that the penny was an Epsom Downs penny, or that it was an ordinary Mint penny, equally likely to fall heads or tails, or if we knew that it was a penny three times more likely to fall heads than tails,—then if 70 heads had been observed in 100 trials, we could say what were the respective probabilities that the penny was an ordinary penny or an Epsom Downs penny.

The criticisms which had been made of the so-called theory of inverse probability and of Bernoulli's theorem had always rested not on the accuracy of the theory itself, but on the correctness of its application, because in most cases these things were not known *a priori* at all.

Given the actual probability in the universe, it was possible to make probability statements about the sort of thing one was likely to get in a sample. These probability statements usually provided a measure for the probability that a certain inequality should be true. Referring to Equation No. 4 in Dr. Neyman's Appendix I, in that equation x was something that belonged to the observed sample, and it was imagined there that we knew for the moment the particular property of the universe.

His own criticism of that particular equation, and of the whole structure placed thereon, was that it was nothing new; it was not a departure from the various earlier attempts that had been made. What was actually done was this: A particular θ was chosen; an inequality was written down, and more values of θ were chosen. In each of these inequalities, ε, which occurred on the right-hand side of the equation, occurred in the definition of inequality on the left side, and had been left out of the equation by Dr. Neyman. When all the points which arose from all the inequalities which could be got by considering all permissible values of θ had been marked, it would be found that we were no further than the man who said, "Let us suppose that the *a priori* probability in the universe is distributed in a particular way. Let us suppose that it fulfils a

certain law." Some had tried to avoid the difficulties by saying, "We can get general results not by assuming that the *a priori* probability satisfies a certain law, but merely by assuming that it is continuous," which meant that if the chance of a penny being exactly symmetrical were so and so, the chance of the penny being nearly symmetrical would be nearly that. As a matter of fact that assumption was equivalent to the absence of gaps among the points on Dr. Neyman's curves; others had tried to follow out the consequences of assuming that the *a priori* probability was continuous near a certain point. All these attempts were rather beating about the bush because the problem was incorrectly stated. When we said that if the probability in the universe were p, then the probability of a certain sample would be x, we were specifying the probability of a certain inequality. It was a matter of elementary algebra to start from that and to say that the inequality so specified, which said that X must be between certain limits in terms of θ and ϵ, led to another inequality which said that given x, θ must lie within certain limits also dependent upon x and ϵ, and that there was a probability for that. The philosophical idea at the bottom of that was rather difficult because we were not now speaking of a probability, but of the probability of a probability. We measured the probability of the truth of the statement that a certain inequality had a particular probability. A hundred years ago mathematicians tried to sum an infinite number of terms in a series, and talked about the ratio of two qualities which ultimately vanished. They happened to be good mathematicians and to have a very sound intuition, and most of their results were correct and had ultimately survived. We had learned that they were occasionally led into a morass, and we said, "Mathematicians cannot perform an infinite number of operations, but can make precise statements about certain inequalities," and it was time that people recognised that while certain probabilities could not be evaluated, correct statements of type $P(P^{(x)})$ could usefully be made. It was possible to go on to higher things, and talk about probabilities of the third or higher orders and still remain in the region of the old subject. The principle to follow was that entities should not be multiplied beyond necessity. This was an argument against belts of confidence and so on, if, as a matter of fact, they only expressed probabilities of statements of the same kind as those which had been made in the past. Dr. Isserlis felt that perhaps he had been wrong, and that he ought to have followed the lead set by Professor Bowley and not trench on mathematics. To try and state mathematics either without chalk or with a minimum of chalk was perhaps a hopeless task. He apologized if he had made himself in any way unintelligible; if something he had meant to say had emerged, he must be satisfied.

Professor Fisher said that the problem of sampling played an important part in Agricultural Research. It was, indeed, in Experimental Agriculture that an adequate technique, bringing out the different aspects of the sampling problem, and displaying comprehensively exactly how these different aspects were interrelated, was

first developed. In the luminous account which Dr. Neyman had given of the sampling technique, as applied to economic researches, which had itself, perhaps, been influenced by his personal experience in Agricultural Science, one of the features which had interested him most had been the parallelism between the processes he advocated (and the reasons he gave for them) on the one hand, with the corresponding processes and reasons which had been developed by agricultural research workers in this country.

His own contact with the subject had been gained at Rothamsted, where he had the pleasure of collaborating with a succession of brilliant plant physiologists, under whom, and especially under Drs. Maskell and Clapham, the technique was gradually perfected. As in agricultural sampling theoretical considerations were at their simplest, the logical connection between the means employed and the inferences which might validly be drawn were conspicuously clear; it might thus be useful if he gave an outline of the hierachy of five successive subdivisions used in the sampling of an agricultural experiment. Exactly the same problems discussed by Dr. Neyman could be simply illustrated in this manner.

The smallest unit that need be considered, the unit of *measurement*, as it might be called, consisted, in the case of a cereal crop, of, perhaps, 10 inches or 25 centimetres measured along a drill row. Again, it might consist of a single plant, as with potatoes or sugar beet. For simplicity he would adhere to the cereal crop. A number of units of measurements, usually four, fixed in relative position, but not necessarily adjacent, constituted a *sampling unit*, which would, therefore, contain in all one metre length of drill row, taken, however, in practice, from four different rows. Since the parts of a sampling unit were fixed in a relative position, the positions of all were determined simultaneously by a single act of random sampling, *i.e.* by the choice, by a physically random process, of the particular sampling unit used from among all those available in the *sampling area*. Two or more sampling units were obtained in this way from each sampling area, each being located independently by a fresh act of randomization. It was essential that there should be at least two independently located sampling units in each sampling area, since it was from the differences between these, or the variances among them, if they were more than two, that the error of sampling was estimated. The variance among the units of measurement within the same sampling unit served a different and subsidiary purpose. It was essential to the study of what structure or size the sampling unit should have, and by analysing the variance within and among sampling units, one could ensure that the sampling units were so chosen as to give the maximum precision in return for the labour expended. But once their size and structure were chosen, this analysis could throw no further light on the interpretation of the experimental results. The error of random sampling, on the other hand, should be ascertained with high precision from every experiment to which the sampling method was applied, for on it one relied for judging of the *number* of sampling units which could with advantage be taken from the growing crop. A usual and convenient number was 32 for each experi-

mental plot. The plot would, therefore, either constitute a single sampling area yielding 32 sample units, or, perhaps, be subdivided into quarters each yielding 8, or at most into sixteenths each yielding 2 sampling units.

Before proceeding to the higher members of the hierachy, it might be useful to indicate a sociological parallel. The sampling unit might be thought of as a family (or as a house, or as a registration district). The sampling area might be thought of as a stratum of such families when they were stratified with respect, say, to earnings. The plot might be thought of as all the families of a given occupational group in a given area, irrespective of their earnings. Then the subdivision of the agricultural plot into sampling areas played the same part in increasing the precision of the ultimate estimates as the stratification of an occupational group according to their earnings. They were, however, ultimately concerned to compare the agricultural plot with other plots which had received different agricultural treatments, just as one might be concerned to compare the morbidity of an occupational group with that of other occupational groups. The sampling was not an essential part of this comparison, but only a convenient means of measurement, which one was concerned, in the first place, to make sufficiently precise.

In an agricultural experiment designed to compare, say 6 different treatments, 48 plots might be assigned to the experiment, and, after dividing the experimental area into 8 compact blocks, each containing 6 plots, these 6 plots should be assigned, strictly at random, to the 6 experimental treatments. This process of experimental randomization could not, unfortunately, be imitated in sociological enquiries. If it could, more than was known would certainly be known about cause and effect in human affairs. But within this limitation the experiment was strictly parallel to one involving a comparison of 6 occupational groups in, say, 8 different towns. In a well-designed experiment, however, the mathematics were simplified, and all anxiety was avoided in respect to different systems of weighting. Dr. Neyman advocated, wisely, in his opinion, the system which he ascribed to Markoff, though this was in essence the system of Gauss. It must be remembered that if the variances from the different populations were not, on a plausible expectation, to be considered equal, one seldom had prior knowledge or experimental evidence sufficient to make the P_i of Dr. Neyman's equation (8) properly speaking known numbers. This seemed to Dr. Fisher a real difficulty, if one wished to speak of the method as the best possible, though it was no obstacle if, as reasonable beings, statisticians were content that it should be a good or valid method. The subdivision into blocks made clear the fact that sampling error was not the only kind of error which had to be considered. Ultimately the validity of the equations must depend upon the concordance of the evidence from the different blocks.

It would be expected that he should comment on those applications of inductive logic which constituted so illuminating and refreshing an aspect of the evening's paper. All realized that problems of mathematical logic underlay all inferences from observational material. They were widely conscious, too, that more than 150 years

of disputation between the *pros* and *cons* of inverse probability had left the subject only more befogged by doubt and frustration. Recently, however, some research workers, working in the apparently abstract realms of the theory of estimation, and the logical bases of tests of significance, had become increasingly confident that, when properly stated, rigorously exact, though, of course, *uncertain* inferences might be drawn from observational or experimental data. In a word, the confidence of the advocates of inverse probability could be confirmed, that valid conclusions of the kind sought could, sometimes, be drawn with assurance, while the arbitrary assumptions upon which from the time of Laplace onwards such inferences had been supported could be rejected as unnecessary. The particular aspect of this work, of which Dr. Neyman's paper was a notable illustration, was the deduction of what Dr. Fisher had called fiducial probability. Dr. Neyman did not use this term, which he suggested had been misunderstood, but he used instead the term "confidence coefficient." Dr. Fisher thought Dr. Neyman must be mistaken in thinking the term fiducial probability had led to any misunderstanding; he had not come upon any signs of it in the literature. When Dr. Neyman said "it really cannot be distinguished from the ordinary concept of probability," Dr. Fisher agreed with him; and that seemed to him a reason for calling it a probability rather than a coefficient. He qualified it from the first with the word *fiducial* to show that it was a probability inferred by the fiducial method of reasoning, then unfamiliar, and not by the classical method of *inverse* probability. Dr. Neyman qualified it with the word *confidence*. The meaning was evidently the same, and he did not wish to deny that confidence could be used adjectivally. They were all too familiar with it, as Professor Bowley had reminded them, in the phrase "confidence trick." Still fiducial was, perhaps, on purely formal grounds, the better adjective.

Dr. Neyman, as he had explained, differed from Dr. Fisher in the relative importance he attached to the two stages in which he had attempted to develop a theory of estimation, independently of all assumptions as to probability *a priori*, namely, the earlier approach through the notions of likelihood and quantity of information, as compared with the later development of the notion of fiducial probability. This difference was not entirely one of perspective. Dr. Fisher's own applications of fiducial probability had been severely and deliberately limited. He had hoped, indeed, that the ingenuity of later writers would find means of extending its application to cases about which he was still in doubt, but some limitations seemed to be essential. Those who had followed the earlier parts of the story would have no difficulty in perceiving these, but there might be pitfalls for those who interested themselves only in the later chapters. In particular, he would apply the fiducial argument, or rather would claim unique validity * for its results, only in those cases

* Naturally, no rigorously demonstrable statements, such as these are, can fail to be true. They can, however, only convey the truth to those who apprehend their exact meaning; in the case of fiducial statements based on inefficient estimates this meaning must include a specification of the process of

for which the problem of estimation proper had been completely solved, *i.e.* either when there existed a statistic of the kind called *sufficient*, which in itself contained the whole of the information supplied by the data, or when, though there was no sufficient statistic, yet the whole of the information could be utilized in the form of *ancillary* information. Both these cases were fortunately of common occurrence, but the limitation seemed to be a necessary one, if they were to avoid drawing from the same body of data statements of fiducial probability which were in apparent contradiction.

Dr. Neyman claimed to have generalized the argument of fiducial probability, and he had every reason to be proud of the line of argument he had developed for its perfect clarity. The generalization was a wide and very handsome one, but it had been erected at considerable expense, and it was perhaps as well to count the cost. The first item to which he would call attention was the loss of uniqueness in the result, and the consequent danger of apparently contradictory inferences.

In the second place, Dr. Fisher had limited his application to continuous distributions, hoping, with more confidence in this case, that the limitation might later be removed. Dr. Neyman removed this limitation, but at the expense of replacing inferences that stated the exact value of the fiducial probability by inequalities, which asserted that it was not less than some assigned value. This also was somewhat a wide departure, for it raised the question whether exact statements of probability were really impossible, and if they were, whether the inequality arrived at was really the closest inequality to be derived by a valid argument from the data.

Thirdly, Dr. Neyman proposed to extend the fiducial argument from cases where there was only a single unknown parameter, to cases in which there were several. Here, again, there might be serious difficulties in respect to the mutual consistency of the different inferences to be drawn; for, with a single parameter, it could be shown that all the inferences might be summarized in a single probability distribution for that parameter, and that, for this reason, all were mutually consistent; but it had not yet been shown that when the parameters were more than one any such equivalent frequency distribution could be established.

Dr. Fisher said that here he ought to point out that Dr. Neyman did him too much honour in ascribing to him the establishment of " Student's " distribution. It was " Student " himself who took the really novel step, which had in fact revolutionized the theory of errors. He showed, in the particular case he treated, that it was possible to find a quantity, which was known to them as " Student's t," having a frequency distribution independent of all unknown parameters, and being at the same time expressible as a function of one only of these parameters, together with other quantities directly observable, and,

estimation employed. But this process is known to omit, or suppress, part of the information supplied by the sample. The statements based on inefficient estimates are true, therefore, so long as they are understood not to be the whole truth. Statements based on sufficient estimates are free from this drawback, and may claim a unique validity.

therefore, known with exactitude. That, as it seemed to Dr. Fisher, constituted the real revolution. All that he had added to it was to "studentize" a number of analogous problems, and to exploit the logical advantages of the position to which he showed the way. It was the more essential that he should make this clear since "Student" was himself far too modest a man to claim, perhaps even to believe, how much he had done for the advance of statistical theory.

The criticism as to the mutual consistency of the different possible inferences did not affect the value of Dr. Neyman's advice on the sampling problem. He stressed it here only because it was just this question of consistency which had led a succession of mathematical writers to reject the theory of inverse probability, and he had no wish to see fiducial probability follow the same course.

The following contribution was received from PROFESSOR OSKAR ANDERSON after the Meeting :—

Dr. Neyman has referred to the enquiry into farming conditions in Bulgaria. As I am responsible for the scientific method applied in it and for the main lines of its organization (the technical execution rested with Dr. Stefanoff, and, of course, with the Chief of the Bulgarian Statistical Office, Dr. Kiranoff), I would like to be permitted to make a few remarks.

The process of fixing the villages to be examined was indeed, in the case of the Bulgarian Enquiry, a difficult one and took a relatively long time. Our opinion was that, for our type of selection, *this* is the central problem and the key of the whole work. We began by picking out all villages in which some special form of farming was highly developed (such as tobacco, roses, rice, silkworms, grape vines, etc.). These formed 13 groups containing about 1,000 villages. The remaining 4,000 villages of Bulgaria, forming a much more homogeneous mass, were divided into 5 climatic regions, and each of these in its turn into 3 physical regions—villages on the mountain slopes, villages in the foothills and villages of the plain. The uniformity of the composition of each group was checked. We then chose for our sample some villages of each group in such a manner that they contained about *one-fiftieth* of the farms of the group, which in their turn gave, for the census data of 1927, the same distribution of the size of estates (in 15 categories !) as that of the whole group from which they were taken. This preliminary choice was then carefully checked by comparing the means of the characteristics of the sample with the means for the whole of Bulgaria *for all data provided by the census of* 1927. It was found, after some corrections in the first choice, that in at least 18 directions out of 19 the agreement became very satisfactory *) This is confirmed by 19 tables and diagrams in the Bulgarian *Bulletin de Statistique*, Number 8, 1934, a copy of which is now in the Library of this society. We tried also to reach a more or

* The only exception is the area of artificial meadow—entirely a *quantité négligeable* in Bulgaria. Some slight discrepancies in the area of vineyards can be easily explained by the existence of vineyards in urban areas (Varna, Plevna, etc.).

less equal distribution of the selected villages on the territory of Bulgaria.

Dr. Neyman finds that there is only one detail in our Enquiry about which he is not certain whether it is justifiable: viz. the purposive manner of selecting the villages from the above 28 "strata," instead of purely random sampling. He thinks "That the variability of farms and villages is also a character of their population which may be of interest." This character, however, would be biased in the sample.

My reply is:—

(1) Our chief object was to determine a great number of general means for the whole of Bulgaria (of 1934) and *not* to measure the variability of any characteristic, and

(2) It can be shown that, owing to (*a*) the very small scale of Bulgarian farming, (*b*) the uniformity of each of the 28 groups, and (*c*) the relatively large number of villages in the selection (100), and (*d*) the approximately linear character of the correlation between the size of farms, and most of their other characteristics—the variability of the farms would not be sensibly biased in a sample containing in any case more than 18,000 farms; and as to the variability of the villages, it is already well known to us.

Without going into mathematics, let us consider the following simple example. Let our problem be to determine the average length of the *right* arm of those present in the Annual Meeting of the Society. Leaving stratification on one side as an unnecessary complication of our example, the method of random sampling would consist in a selection by chance of, say, 10 members of the audience, measuring their right arms and calculating the mean. But suppose that we know *beforehand* that the mean length of the *left* arms of the whole group present is, say, 28 inches: if then we again select 10 members of the audience whose left arm is of or about this length, then Dr. Neyman will agree that it is most probable that the new mean length of *right* arms (estimated by a "purposive selection") will be more accurate than the former mean (estimated on a purely random sample). But, if our selection were based, *e.g.* on the mean colour of ties worn by members of the audience, the result would be much less useful. In the Bulgarian Enquiry, I think, the analogy was much closer to control by length of left arm than to control by colour of ties.

Why did we use purposive selection and not random sampling, which I also personally prefer? The reasons are very simple—time and money. Given our very detailed programme, the short time and the limited qualified personnel available in Bulgaria, it was by no means possible to extend the enquiry over more than 100 villages. The work had to be finished in one month, before the beginning of farming operations, and, of course, we could not drive our relatively small staff over mountains, through forests, water and snow in early spring and on Bulgarian roads to, say, 1,000 villages, which would be necessary in the case of a purely random sampling of them, even if stratified.

The Bulgarian Statistical Office hopes to publish the results of our

enquiry at the end of this year. The report will contain a theoretical introduction and all possible controls on the general lines indicated by the formulæ of Prof. Bowley, and also of Dr. Neyman.

To conclude: I fully agree with Dr. Neyman that with random sampling the statistician is on firmer ground and feels more confidence than with purposive selection. But I differ from him in thinking that there *are* occasions where the latter is both more economical and more exact. The conditions for a good purposive selection are, firstly, that the object of the enquiry should possess some qualities similar to those of the Bulgarian or Danish * example, and secondly, that all the circumstances of the enquiry must be very carefully planned and checked. I agree that both are far from being always possible; and in any case, in my opinion, the first thing which is needed for a successful issue is by no means "good luck," as Dr. Neyman says, but a perfect knowledge of the object and sound statistical reasoning.

Dr. Neyman, in reply: I am most grateful to all those who have taken part in the discussion of my paper. Extensive discussion is very useful, not only to listeners and readers, but also to the author, because it shows him what is properly done in his paper and where he is at fault.

The present discussion has shown, I think, (i) that my criticism against the method of purposive selection was sufficiently convincing, and (ii) that the sections concerned with the confidence intervals and the problem of estimation were not. Out of the four eminent statisticians who have honoured my paper by discussing it, there was only one who defended the method of purposive selection. And then this was not a defence of the method itself, but rather of a separate inquiry in which a mixture of the two methods of stratified sampling by groups, and of purposive selection was used. Professor Bowley stated that he distrusted the method of purposive selection, even in 1925, when he prepared his Report to the International Statistical Institute.

Therefore, as far as the main subject of my paper is concerned, I have to argue only with Professor Anderson, and I am glad that the argument will concern only details, not my main thesis. In fact, his statement that, "with random sampling, the statistician is on firmer ground and feels more confidence than with purposive selection," shows that we are in perfect agreement on the main point. Our agreement extends even further than Professor Anderson seems to think, as I agree with him that in *certain* cases the method of purposive selection may be applied with great success. I even suggested a class of problems in which this is the case (see the last paragraph of the main text of the paper, page 589). But we are relying on good luck if we apply the method without sufficient evidence of its validity.

It is a very interesting fact that, in Bulgarian conditions, the regressions of many characters of farms on their size are linear. In

* See Adolph Jensen, "Purposive Selection," *Journ. Roy. Stat. Soc.*, Vol. XCI, pp. 541–547.

Poland we have found * that many regressions, for instance the regression of the gross income on the size and on the outlay, cannot be represented by a plane, and that the partial regression of the income on the outlay depends very much upon the size of the farms. Probably the general farming conditions in the two countries are very different. The application of Professor Anderson's method in Poland would lead to biased estimates of the means sought.

With regard to the problem of determining the average length of the right arm of those present at the meeting, which has been quoted by Professor Anderson as an instance in which the method of purposive selection should be applied, I should like to notice that it is a special case of the class of the problems indicated in the last sentence of my paper (page 589) as suitable for the application of this method. So here again we are in agreement. I cannot agree, however, that the accuracy of an unbiased estimate will be increased if we purposely select individuals with the length of the left arm approximately equal to its average. The solution of this question follows directly from the formulae, already fully discussed in text-books,† to the effect that the greater accuracy is obtained (i) by minimising the difference between the sample and the population means of the control, and (ii) by maximizing the sample standard deviation of the control.

* See: (1) W. Pytkowski: *The Dependence of the Income in small Farms upon their Area, the Outlay and the Capital invested in Cows.* Bibljoteka Pulawska, Warsaw, 1932.

(2) K. Iwaszkiewicz: "*La rentabilité de l'étendue, du fonds de roulement et du capital investi en vaches dans les petites exploitations rurales.*" Kwartalnik Statystyczny, t. IX, Fascicule 1, 1932, Warsaw.

(3) M. Iwaszkiewicz: "*Recherches statistiques sur la rentabilité des engrais artificiels dans les petites exploitations rurales.*" Kwartalnik Statystyczny, t. X, Fascicule 2–3, 1933, Warsaw.

All these publications are to be found in "Statistica"—the collections of the papers prepared in the Biometric Laboratory, Nencki Institute and in the Statistical Laboratory, Central College of Agriculture, Warsaw.

† See, for instance (1) M. Ezekiel: "*Methods of Correlation Analysis,*" New York, 1930, pp. 252—255.

(2) R. A. Fisher: "*Statistical Methods for Research Workers,*" London, 1932, pp. 115–117.

Finally, the solution may be obtained from a suitable adjustment of formulae (36), (37) and (39) in the Note III of the Appendix, p. 603.

The mentioned formulae are :—

$$Y' = \bar{y} + a(X - \bar{x})$$

and

$$\mu^2 = \frac{\sigma_y^2(1 - r^2)}{n - 2}\left(1 + \frac{(X - \bar{x})^2}{\sigma_x^2}\right)$$

where $\bar{x}$ and $\bar{y}$ mean the sample means of the two variables, σ_x^2 and σ_y^2 the sample variances, X the known population mean of x, Y'—the estimate of the population mean of y, μ^2—the estimate of the variance of the same, a—the sample regression coefficient of y on x and r—the sample correlation coefficient. As the difference $| X - \bar{x} |$ is practically never zero, in order to attain the greatest accuracy of the estimate Y', it is advisable to try to minimize $| X - \bar{x} |$ keeping σ_x as large as possible. In order to do so, it is necessary to include in the sample individuals both with very large values of x and with very small ones. Then the difference $| X - \bar{x} |$ will be small and σ_x large, and μ^2 will approach its minimum value $\frac{\sigma_y^2(1 - r^2)}{n - 2}$.

I did not expect that the sections of my paper dealing with the new form of the problem of estimation would play so large a part in the discussion. It is at least gratifying that the criticisms were so divergent that one of the speakers could say that everything in it is doubtful, and another that it is nothing new. I considered it necessary to include these sections in my paper, as otherwise it would not be complete, and it would be justifiable to ask why I am not troubling to consider the problems which Professor Bowley termed the inverse problems.

Detailed comments on all questions raised in the discussion on the confidence intervals would require too much space. In fact, to clear up the matter entirely, a separate publication is needed. As this is in preparation, I shall limit myself only to one or two remarks, which may clear away certain obvious misunderstandings.

It has been suggested in the discussion that I used the term " confidence coefficient " *instead* of the term " fiducial probability." This is certainly a misunderstanding. The term confidence coefficient is not synonymous to the term probability. It means an arbitrarily chosen value of the probability of our being right when applying a certain rule of behaviour. The relation of the concept of the confidence coefficient to that of probability may be compared to the relation between the concepts of the " price " and " money " (this, if we accept the definition of the " price " as " a certain amount of money which has been fixed by the merchant . . ."). Perhaps a still better comparison is provided by the terms " rate of interest " and " money." The analogy here is less superficial than one would expect. Banks are working at a certain rate of interest, which is being fixed once for a longer period, and just this constancy led to the introduction of the term " rate of interest." The validity of probability statements in the new form of the problem of estimation, which has been here so extensively discussed, depends on the permanent use of a system of confidence intervals. This system as a whole (not separate intervals) corresponds to a fixed probability that our predictions are correct, and certainly there is a definite advantage in having a special term to denote this value of the probability. It would allow us, for example, to use the convenient expressions like the following :—the seed-testing station in X is working with the confidence coefficient ·95, etc.

Another important misunderstanding, which I think it useful to clear up now, is contained in the following remarks of Professor Bowley concerning the theory of the confidence intervals :— " (*a*) Does it really lead us towards what we need—the chance that, in the universe which we are sampling, the proportion is within these certain limits ? I think it does not; (*b*) I think we are in the position of knowing that *either* an improbable event has occurred, *or* the proportion in the population is within the limits."

I have marked the two sentences with letters (*a*) and (*b*) as I shall have to comment on them separately.

The sentence (*a*) contains the statement of the problem of estimation in the form of Bayes. Simple algebra shows that the solution of this problem *must* depend upon the probability law *a priori*.

Therefore, if all we need consists in "the chance that, in the universe which we are sampling, the proportion is within given limits," we certainly cannot go any further than it is already known.

In so far as we keep to the old form of the problem, any further progress is impossible. It would be possible only if the previous writers on the subject had been wrong. The present progress is connected with the fact that, *instead* of the mathematical problem stated in the sentence (*a*), we are solving some other mathematical problem, say (α), which (i) has a solution independent of any arbitrary assumptions concerning the probability law *a priori*, and (ii) may form a basis for the practical work of a statistician concerned with problems of estimation.

Now what is the difference between the problems (*a*) and (α)? Both of them are dealing with probabilities, but these probabilities apply to different events. In the problem (*a*) we ask about the probability that a character of the sampled population lies within certain limits, while in the problem (α), we are interested in the probability of committing an error when applying constantly a certain rule of behaviour. The former probability proved to be dependent on the probabilities *a priori*. On the other hand, it was possible to invent such systems of statements about the values of population characters that the probability of being wrong in these statements is a fixed one, whatever the probabilities *a priori* (the law of which may be continuous or not, without restriction). This circumstance, that in the problem of confidence intervals the probability statements concern the results of our behaviour, not the populations and that they relate to this given rule of behaviour, not to the properties of samples to which this rule is being applied, is very important.

Next, let me remark on the sentence (*b*) in the criticism of Professor Bowley. I shall simplify the position, assuming that we want to determine the proportion, p, of black balls in a bag, and for this purpose we intend to draw as many as $n = 3$ balls (with replacement—for simplicity). The number of black balls in the sample will be denoted by X, which may have the values 0, 1, 2, 3. The sentence (*b*) of Professor Bowley assumes that we actually have the sample drawn and that, for instance, $X = 3$. Having got so far, we certainly cannot tell much about the probability of p having any definite value. In fact, all we can say is that either an improbable fact occurred, or p has a rather large value. But what is done in the method of confidence intervals is different. We start our reasoning, as it were, before drawing the sample. Or again, we assume that this is not the only one sample with which we shall have to deal. We notice that the sample may provide us only with one out of the four possibilities $X_1 = 0$, $X_2 = 1$, $X_3 = 2$ and $X_4 = 3$. Having noticed this, we fix a rule as follows :—

If in the sample which we shall draw, X will have the value

If $X = 0$, then we shall state that $0 \leq p \leq \pi_1''$
$X = 1$, ,, ,, ,, $\pi_2' \leq p \leq \pi_2''$
$X = 2$, ,, ,, ,, $\pi_3' \leq p \leq \pi_3''$
$X = 3$, ,, ,, ,, $\pi_4' \leq p \leq 1$

We are aware that the statement which we shall make, in applying this rule to the result of actual sampling, may be wrong or may be true. We calculate the probability, P, that the statement will be a true one, and try to arrange the system of values of the π's so as to have $P \geq \cdot 95$, whatever the probability law *a priori*.

In this way, having a sample and knowing that $X = 3$, we cannot tell any more about the value of p than Professor Bowley has stated. On the other hand, making statements following the rules set out above, we know something important about the results of these statements : the probability that we shall be wrong is then $\leq \cdot 05$.

The last misunderstanding I should like to clear concerns the recognition of the merits of the work of " Student." I certainly recognize and appreciate it very much. If I call a distribution " Student's " distribution, it means clearly that I attribute its discovery to " Student," and not to anybody else. This does not prevent me from recognizing and appreciating the work of Professor Fisher concerning the same distribution.

I am very grateful to all present at the meeting for the kind reception of my paper.