ON THE MATHEMATICS OF THE REPRESENTATIVE METHOD OF SAMPLING[1]

By Allen T. Craig

1. Introduction. This paper is designed to present certain topics in mathematical statistics which find application in some of the problems that arise in what has been termed the representative method of sampling.

For descriptive purposes, it seems convenient to consider two aspects of the representative method. The first of these may be called the method of *purposive selection*. This method can be roughly characterized by saying that it is the method employed when the samples are chosen in such a way that each sample will possess one or more characters, say certain averages, which are identical with the corresponding characters in the population from which the samples are drawn. The mathematical conditions which underlie this method are rather stringent, and both theoretical and practical investigations seem to have proved that in general no great amount of confidence can be placed in the results obtained.

The second aspect of the representative method has been styled the method of *random sampling*. This method can take either of two forms which we may call the method of *unrestricted random sampling* and *stratified random sampling*. The first of these is the classical method of procedure. That is, a sample is drawn at random from a given population and on the basis of these data inferences are made concerning the nature of the population. On the other hand, when the method of stratified random sampling is used, the population is first separated into a large number of parts, called strata, and the sample consists of an equally large number of "partial samples," each partial sample being drawn from a different stratum. It appears, both from theoretical and practical results, that this method of stratified random sampling enjoys many advantages not shared by the other methods.

We now turn to the main purpose of this paper, namely that of enumerating some of the theorems and methods of mathematical statistics which serve useful purposes in this theory. Discussion of how these theorems find application in the method itself has been reserved for other participants on this program.

2. Estimates. From our preliminary remarks, it is apparent that the representative method is much concerned with the problem of estimating certain

---

[1] Presented, at the invitation of the program committee, to a joint session of the Institute of Mathematical Statistics and the American Statistical Association on December 29, 1938.

unknown *parameters* of a statistical population. On this account, we first consider the problem of estimates.

Consider a population with arithmetic mean m and standard deviation σ. Let $x_1, x_2, \cdots, x_n$, be n independent items drawn from this population and let $c_1, c_2, \cdots, c_n$ be any finite real constants, not all zero to avoid the trivial case. Write $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$. Then the expected or arithmetic mean value of y is

$$\bar{y} = E(y) = m(c_1 + c_2 + \cdots + c_n),$$

and the variance of y is

$$\sigma_y^2 = E\{(y - \bar{y})^2\} = \sigma^2(c_1^2 + \cdots + c_n^2).$$

Suppose we inquire into the probability that y will have a value which is within a preassigned ϵ of its expected value. To this end, let C be the numerical value of the numerically greatest of the set $c_1, \cdots, c_n$, so that $\sigma_y^2 \leq n\sigma^2C^2$. Then by Tchebycheff's inequality p, the probability that $|y - \bar{y}| < \epsilon$, where ϵ is an arbitrarily small positive number, is such that

$$p \geq 1 - \frac{\sigma_y^2}{\epsilon^2},$$

or

$$p \geq 1 - \frac{n\sigma^2C^2}{\epsilon^2}.$$

In general, this inequality will have little interest. But if C is of the form $M/n^{\frac{1+\delta}{2}}$, M independent of n, $\delta > 0$, then $p \geq 1 - \frac{\sigma^2M^2}{n^\delta\epsilon^2}$ and by increasing n the right member can be made as near to one as we please. This means then that if we have a population with a finite variance and if we construct a linear function of the observations with coefficients of the nature indicated, we can, by increasing the size of the sample, make the probability approach one that the linear function will have a value arbitrarily close to its expected value.[2]

Now suppose that instead of constructing an arbitrary linear function we attempt to construct a function which will be an estimate of some particular parameter of the population. If the estimate is to be most serviceable, we should like to be able, by governing the size of the sample, to be as certain as we like that the estimate will have a value arbitrarily near that of the parameter. The preceding discussion shows that we can best achieve this by requiring that the expected value of the estimate be equal to the parameter sought. An estimate such as that just described is frequently called an *unbiased* estimate. The use of such estimates in statistical problems makes it possible to avoid systematic errors in estimating parameters. In general, unique unbiased estimates of a parameter do not exist. For example, the arithmetic mean m of

[2] Under these conditions, the function of the observations is said to converge stochastically to its expected value.

the population can be estimated from the sample $x_1, \cdots, x_n$ by any one of a large number of unbiased estimates such as $(x_1 + x_2 + \cdots + x_n)/n$, $(x_1 + x_n)/2$, x_4, and so on without limit. Thus it becomes necessary to make a choice of the unbiased estimate to be used. An appropriate criterion is that the unbiased estimate whose distribution has the smallest variance is the best to use. The reason for this can be seen by examining the preceding formula $p \geq 1 - \frac{\sigma_y^2}{\epsilon^2}$. For if y_1 and y_2 are two unbiased estimates of the same parameter and if $\sigma_{y_1}^2 < \sigma_{y_2}^2$, then in $p_1 \geq 1 - \frac{\sigma_{y_1}^2}{\epsilon^2}$ and $p_2 \geq 1 - \frac{\sigma_{y_2}^2}{\epsilon^2}$ we see that $1 - \frac{\sigma_{y_1}^2}{\epsilon^2}$ is more nearly equal to one than is $1 - \frac{\sigma_{y_2}^2}{\epsilon^2}$. Because of this fact we prefer, at least in most problems, to use y_1 rather than y_2 as an estimate of the unknown parameter. An unbiased estimate whose sampling variance is a minimum is sometimes called a *best estimate.*[3] It should not be inferred that the word "best" has any implications other than those stated explicitly in the definition.

The question very naturally arises as to whether we can determine these best estimates in particular cases. In general we can not determine them, but under certain conditions we can find best estimates if we are dealing with linear functions of the observations. A method and the conditions are set forth in an important theorem due to Markoff. We now consider his method.

3. Markoff's Method. Let there be given n statistical populations with arithmetic means $m_1, m_2, \cdots, m_n$ and standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_n$ respectively. We assume that no correlation exists between any of the populations. Furthermore, suppose that each of the n arithmetic means can be expressed linearly in terms of k unknown, but unique, parameters, say $z_1, z_2, \cdots, z_k$. Thus

$$
\begin{aligned}
m_1 &= a_{11}z_1 + a_{12}z_2 + \cdots + a_{1k}z_k \\
m_2 &= a_{21}z_1 + a_{22}z_2 + \cdots + a_{2k}z_k \\
&\vdots \\
m_n &= a_{n1}z_1 + a_{n2}z_2 + \cdots + a_{nk}z_k,
\end{aligned}
\tag{1}
$$

where the a's are known constants. Likewise, let T be a parameter which is expressible linearly in terms of the same k unknown parameters, say $T = b_1z_1 + b_2z_2 + \cdots + b_kz_k$, where the b's are given constants. We draw a sample of n independent items, $x_1, x_2, \cdots, x_n$, in which one item is drawn from each

[3] An estimate of a parameter which converges stochastically (cf. footnote (2)) to that parameter is called a *consistent* estimate of the parameter. If a consistent estimate has a distribution which is normal for large samples and if the variance of that distribution is smaller than the variance of any other consistent estimate which also has a normal distribution for large samples, then the estimate is called *efficient*. It should be observed that our definition of best estimate requires an unbiased estimate, whereas consistent and efficient estimates may be biased.

of the n populations. On the basis of this sample we seek to determine a set of numbers $\lambda_1, \lambda_2, \cdots, \lambda_n$ such that $T' = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$ is the best estimate of T.

Before attempting to find the solution, if one exists, let us first examine the mathematical implications of the problem. In the first place, in order that parameters $z_1, \cdots, z_k$ may exist, it is necessary and sufficient that the matrices A and B, where

$$A = \begin{Vmatrix} a_{11} & a_{12} \cdots a_{1k} \\ a_{21} & a_{22} \cdots a_{2k} \\ \vdots & \\ a_{n1} & a_{n2} \cdots a_{nk} \end{Vmatrix} \text{ and } B = \begin{Vmatrix} a_{11} & a_{12} \cdots a_{1k} & m_1 \\ a_{21} & a_{22} \cdots a_{2k} & m_2 \\ \vdots & & \\ a_{n1} & a_{n2} \cdots a_{nk} & m_n \end{Vmatrix}$$

have the same rank. Thus we require that A and B have the common rank R. This being satisfied, we note further that if $k > n$, there will be infinitely many values of the z's which will satisfy the equations (1). Thus we require in addition that $k \leq n$. Finally, we note that if the common rank R is less than k, there will be infinitely many values of the z's which will satisfy the system (1). Hence we must have $R = k \leq n$.

We now turn to a consideration of the solution of the problem. Whatever the values of the λ's, we have for the mean value and the variance of T'

$$\begin{aligned} E(T') &= \lambda_1 m_1 + \cdots + \lambda_n m_n \\ &= \lambda_1 \Sigma a_{1j} z_j + \cdots + \lambda_n \Sigma a_{nj} z_j , \end{aligned}$$

and

$$\sigma_{T'}^2 = \lambda_1^2 \sigma_1^2 + \cdots + \lambda_n^2 \sigma_n^2 ,$$

respectively. Since $E(T')$ must equal T as a part of the condition for a best estimate, then

$$\lambda_1 \Sigma a_{1j} z_j + \cdots + \lambda_n \Sigma a_{nj} z_j = b_1 z_1 + \cdots + b_k z_k$$

identically in the z's. That is, the coefficients of $z_1, \cdots, z_k$ in the left member must equal the corresponding coefficient in the right member. Accordingly,

$$\begin{aligned} a_{11}\lambda_1 + a_{21}\lambda_2 + \cdots + a_{n1}\lambda_n &= b_1 \\ a_{12}\lambda_1 + a_{22}\lambda_2 + \cdots + a_{n2}\lambda_n &= b_2 \\ \vdots \qquad & \\ a_{1k}\lambda_1 + a_{2k}\lambda_1 + \cdots + a_{nk}\lambda_n &= b_k . \end{aligned} \tag{2}$$

If these equations are to have solutions for $\lambda_1, \cdots, \lambda_n$, we must make the additional assumption that the matrix C, where

$$C = \begin{Vmatrix} a_{11} & a_{21} \cdots a_{n1} & b_1 \\ a_{12} & a_{22} \cdots a_{n2} & b_2 \\ \vdots & & \\ a_{1k} & a_{2k} \cdots a_{nk} & b_k \end{Vmatrix}$$

has the same rank as the matrix of the coefficients, namely R. If this condition is satisfied we can write equations (2) in the form

$$
\begin{aligned}
&a_{11}\lambda_1 + \cdots + a_{k1}\lambda_k = b_1 - a_{k+1,1}\lambda_{k+1} - \cdots - a_{n1}\lambda_n \\
(3) \qquad &\vdots \\
&a_{1k}\lambda_1 + \cdots + a_{kk}\lambda_k = b_k - a_{k+1,k}\lambda_{k+1} - \cdots - a_{nk}\lambda_n
\end{aligned}
$$

and solve for $\lambda_1, \cdots, \lambda_k$ in terms of the a's, the b's, and $\lambda_{k+1}, \cdots, \lambda_n$. Here, without any essential loss of generality, we take the non-vanishing k-rowed determinant to be that of the coefficients of $\lambda_1, \cdots, \lambda_k$ in equations (2). Thus for arbitrarily assigned values of $\lambda_{k+1}, \cdots, \lambda_n$, we can compute the values of $\lambda_1, \cdots, \lambda_k$ and these n values of the λ's will give us a T' which is an unbiased estimate of T. That there will be, in general, an unlimited number of sets of values of the λ's is in keeping with our previous observation that unique unbiased estimates usually do not exist.

The next part of the problem will consist in determining which, if any, of the above sets of λ's will make $\sigma^2_{T'}$ a minimum. We recall that $\sigma^2_{T'} = \lambda_1^2\sigma_1^2 + \cdots + \lambda_n^2\sigma_n^2$. In $\sigma^2_{T'}$ let us replace $\lambda_1, \cdots, \lambda_k$ by their values (in terms of $\lambda_{k+1}, \cdots, \lambda_n$) which we obtained by solving the system (3). Then $\sigma^2_{T'}$ will be expressed in terms of $\sigma_1, \cdots, \sigma_n$, the a's, the b's, and $\lambda_{k+1}, \cdots, \lambda_n$. We next take the partial derivative of $\sigma^2_{T'}$ with respect to each of $\lambda_{k+1}, \cdots, \lambda_n$. On equating these partial derivatives to zero we will have a system of $n - k$ linear equations in the $n - k$ unknowns $\lambda_{k+1}, \cdots, \lambda_n$. If these equations yield unique values for $\lambda_{k+1}, \cdots, \lambda_n$, they will in turn determine unique values of $\lambda_1, \cdots, \lambda_k$. This gives us a unique set of λ's such that at one and the same time

$$E(T') = T \text{ and } \sigma^2_{T'} \text{ is a minimum.}$$

The procedure which we have just outlined is most tedious to carry out in a particular case. Because of the insight of Markoff, a much better scheme is available for finding the best estimate of T. Consider the function of $z_1, \cdots, z_k$,

$$
\begin{aligned}
F(z_1, \cdots, z_k) &= \sum \left(\frac{x_j - m_j}{\sigma_j}\right)^2 \\
&= \sum \left(\frac{x_j - a_{j1}z_1 - \cdots - a_{jk}z_k}{\sigma_j}\right)^2.
\end{aligned}
$$

Evaluate $\frac{\partial F}{\partial z_1}, \cdots, \frac{\partial F}{\partial z_k}$ and equate these partial derivatives to zero. This yields the following system of k linear equations in the k unknowns $z_1, \cdots, z_k$.

$$
\begin{aligned}
&z_1 \sum \frac{a_{j1}^2}{\sigma_j^2} + \cdots + z_k \sum \frac{a_{j1}a_{jk}}{\sigma_j^2} = \sum \frac{a_{j1}x_j}{\sigma_j^2} \\
(4) \qquad &\vdots \\
&z_1 \sum \frac{a_{j1}a_{jk}}{\sigma_j^2} + \cdots + z_k \sum \frac{a_{jk}^2}{\sigma_j^2} = \sum \frac{a_{jk}x_j}{\sigma_j^2}.
\end{aligned}
$$

If the system (4) yields unique values for the z's, these values, when substituted in T, yield exactly the same estimate of T as was found by substituting for the λ's in T'.

Perhaps an illustration will make this clearer. Suppose we have $n = 2$ populations and that the means m_1 and m_2 are expressible linearly in terms of $k = 1$ parameter z_1. Our equations (1) become

$$\begin{aligned} m_1 &= a_{11} z_1 \\ m_2 &= a_{21} z_1 . \end{aligned} \tag{1'}$$

Similarly, we have $T = b_1 z_1$ and $T' = \lambda_1 x_1 + \lambda_2 x_2$. We first determine the λ's such that T' is the best estimate of T. In accordance with the preceding steps, equations (2) become

$$a_{11}\lambda_1 + a_{21}\lambda_2 = b_1, \tag{2'}$$

and the system (3) becomes

$$\lambda_1 = \frac{b_1 - a_{21}\lambda_2}{a_{11}}, \qquad a_{11} \neq 0. \tag{3'}$$

Then

$$\begin{aligned} \sigma_{T'}^2 &= \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 \\ &= \left(\frac{b_1 - a_{21}\lambda_2}{a_{11}}\right)^2 \sigma_1^2 + \lambda_2 \sigma_2^2, \end{aligned}$$

because of (3′). Thus

$$\frac{\partial \sigma_{T'}^2}{\partial \lambda_2} = \frac{-2a_{21}(b_1 - a_{21}\lambda_2)\sigma_1^2}{a_{11}^2} + 2\lambda_2 \sigma_2^2,$$

and for a minimum $\sigma_{T'}^2$ we write $\dfrac{\partial \sigma_{T'}^2}{\partial \lambda_2} = 0$ so that

$$\lambda_2 = \frac{a_{21} b_1 \sigma_1^2}{a_{11}^2 \sigma_2^2 + a_{21}^2 \sigma_1^2}.$$

Since

$$\lambda_1 = (b_1 - a_{21}\lambda_2)/a_{11},$$

then

$$\lambda_1 = \frac{b_1 a_{11} \sigma_2^2}{a_{11}^2 \sigma_2^2 + a_{21}^2 \sigma_1^2}.$$

Our best estimate of T is found from T' and it is

$$T' = \frac{b_1 a_{11} \sigma_2^2 x_1 + b_1 a_{21} \sigma_1^2 x_2}{a_{11}^2 \sigma_2^2 + a_{21}^2 \sigma_1^2}.$$

By Markoff's method we would form the function

$$F(z_1) = \sum \left(\frac{x_j - a_{j1} z_1}{\sigma_j}\right)^2 = \left(\frac{x_1 - a_{11} z_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - a_{21} z_1}{\sigma_2}\right)^2.$$

The system (4) reduces to merely

$$z_1 = \frac{a_{11}\sigma_2^2 x_1 + a_{21}\sigma_1^2 x_2}{a_{11}^2\sigma_2^2 + a_{21}^2\sigma_1^2}. \tag{4'}$$

We substitute this value of z_1 in $T = b_1 z_1$ and obtain

$$T = \frac{b_1 a_{11}\sigma_2^2 x_1 + b_1 a_{21}\sigma_1^2 x_2}{a_{11}^2\sigma_2^2 + a_{21}^2\sigma_1^2},$$

which is the estimated value T' above.

4. Neyman's modification of Markoff's Method. We are indebted to Neyman for a modification and adaptation of the Markoff method so as to make the method applicable to some of the problems of stratified random sampling. One of his examples will best illustrate the method.

Suppose that a given population is divided into n strata. Let the jth stratum contain M_j items and let these items be $u_{j1}, u_{j2}, \cdots, u_{jM_j}$. The mean and the variance of this stratum are then

$$\bar{u}_j = \frac{1}{M_j}\sum_k u_{jk} \quad \text{and} \quad \sigma_j^2 = \frac{1}{M_j}\sum_k (u_{jk} - \bar{u}_j)^2.$$

Let T be the parameter $T = M_1\bar{u}_1 + M_2\bar{u}_2 + \cdots + M_n\bar{u}_n$, so that $\dfrac{T}{M_1 + \cdots + M_n}$, the mean of the population, is expressed as a linear function of the means of the n strata. We draw at random a sample of N items, the sample consisting of n partial samples, one partial sample being drawn from each of the n strata. Suppose there are n_1 items in the partial sample from the first stratum, n_2 from the second, and so on. Thus $n_1 + n_2 + \cdots + n_n = N$ and the entire sample consists of the n partial samples

$$\begin{array}{c} x_{11}, x_{12}, \cdots, x_{1n_1} \\ x_{21}, x_{22}, \cdots, x_{2n_2} \\ x_{n1}, x_{n2}, \cdots, x_{nn_n}. \end{array}$$

From these N data we propose constructing an estimate

$$T' = \lambda_{11}x_{11} + \cdots + \lambda_{1n_1}x_{1n_1} + \cdots + \lambda_{n1}x_{n1} + \cdots + \lambda_{nn_n}x_{nn_n}$$

which will be the best estimate of T. Now the expected value of T' is

$$\begin{aligned} E[T'] = E\Big[\sum_j \sum_k \lambda_{jk} x_{jk}\Big] &= \sum_j \sum_k \lambda_{jk} E(x_{jk}) \\ &= \sum_j \sum_k \lambda_{jk} \bar{u}_j \\ &= \sum_j \bar{u}_j \sum_k \lambda_{jk}, \end{aligned}$$

which, by hypothesis, must equal T. Thus

$$\sum_1^n \bar{u}_j \sum_1^{n_j} \lambda_{jk} = \sum_1^n M_j \bar{u}_j$$

identically in the $\bar{u}$'s. Hence $\Sigma \bar{u}_j(M_j - \Sigma\lambda_{jk}) = 0$ which requires that the coefficients of $\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_n$ must be zero. That is

$$\sum_1^{n_1} \lambda_{1k} = M_1$$

$$\vdots$$

$$\sum_1^{n_n} \lambda_{nk} = M_n.$$

Of course there are infinitely many λ's which will satisfy these equations. But we can eliminate all but one set by imposing the condition that $\sigma^2_{T'}$ shall be a minimum. The algebra of mathematical expectation can be used to show that

$$\sigma^2_{T'} = \sum_1^n \sigma_j^2 \left[\frac{M_j n_j - n_j^2}{M_j - 1} \left(\frac{1}{n_j} \sum \lambda_{jk} \right)^2 + \frac{M_j}{M_j - 1} \sum \left(\lambda_{jk} - \frac{1}{n_j} \sum \lambda_{jk} \right)^2 \right]$$

which will be a minimum when $\sum \left(\lambda_{jk} - \frac{1}{n_j} \sum \lambda_{jk} \right)^2 = 0, j = 1, 2, \cdots, n$. Since this is a sum of real squares, each term in the sum must be zero. Thus, $\lambda_{jk} = \frac{1}{n_j} \sum \lambda_{jk}$. Since $\sum \lambda_{jk}$ must equal M_j in order that $E(T') = T$, then $\lambda_{jk} = \frac{M_j}{n_j}$ which uniquely determines the λ's and hence our best estimate of T'.

It is important to observe that Neyman's adaptation does not assume that the various strata are uncorrelated nor that there are necessarily replacements after each drawing in taking the sample.

5. **Estimation of Ratios.** In certain problems in representative sampling it may be necessary to estimate both the numerator and the denominator of a fraction, say T/U. If T' and U' are linear estimates of T and U then for large samples both T' and U' will be approximately normally distributed in most cases. Further, if T' and U' are correlated, they will usually be approximately normally correlated. Geary has proved that if we write

$$V = \frac{b + T'}{a + U'},$$

where a and b are constants and U' and T' are measured from their expected values, then

$$t = \frac{aV - b}{\sqrt{V^2\sigma^2_{U'} - 2rV\sigma_{T'}\sigma_{U'} + \sigma^2_{T'}}}$$

is approximately normally distributed with mean zero and unit variance provided $a \geq 3\sigma_{U'}$. Here r is the correlation coefficient between T' and U'. For large samples this provides a convenient method of testing the significance of the difference between an observed and a hypothetical ratio of two linear estimates.

6. Fiducial Inference. After an estimate of a parameter has been made, it is usually desirable to make some inference about the true value of the parameter. For many years the concept of probable error was used in this connection. But the use of the probable error involves the assumption that all values of the unknown parameter are equally likely. This assumption is questionable and efforts to avoid making the assumption have led to a theory called *fiducial inference*. This method of statistical inference has broad implications but limitations on our time do not permit our discussing the topic. At the close of this paper, we give certain references to the subject, including some of an expository nature.

7. Conclusion. As stated in the introduction, this paper purports to give an exposition of some of the topics in mathematical statistics which find application in the representative method of sampling. Necessarily considerable selection of material had to be made. We believe, however, that the problem of the best estimate and an appropriate method of obtaining such an estimate are fundamental, and we hope that our exposition has helped to make clear these concepts of mathematical statistics which have proved so useful in the representative method.

REFERENCES

[1] J. Neyman, "On the two different aspects of the representative method," *Journal of the Royal Statistical Society*, vol. 97 (1934) pp. 558–625.

[2] J. Neyman, *Lectures and conferences on mathematical statistics, Graduate School of the U. S. Department of Agriculture*, (1937) pp. 89–105.

[3] J. Neyman, "Contributions to the theory of sampling human populations," *Journal of the American Statistical Association*, vol. 33 (1938) pp. 101–116.

[4] S. S. Wilks, *Statistical inference*, Edwards Brothers (1936–37) pp. 8–11, p. 70, pp. 71–86.

[5] R. C. Geary, "The frequency distribution of the quotient of two normal variates," *Journal of the Royal Statistical Society*, vol. 93 (1930) pp. 442–446.

[6] R. A. Fisher, "Inverse probability," *Proceedings of the Cambridge Philosophical Society*, vol. 26 (1930) p. 528.

[7] H. L. Rietz, "A recent advance in statistical inference," *American Mathematical Monthly*, vol. 45 (1938) pp. 149–158.

The University of Iowa.