141

"THE COEFFICIENT OF RACIAL LIKENESS" AND THE FUTURE OF CRANIOMETRY.

By R. A. Fisher.

Scientific specialisation may be justified as a form of the division of labour. In theory nothing could be more simpler. Of the many things which need doing some require special aptitude or special training. If individuals with the necessary ability can be induced to occupy themselves with these tasks, and to spend a large part of their working life in acquiring experience and proficiency, we may expect that these particular jobs will be better done than would otherwise be possible. The simplicity of the plan is elementary ; but there is one supposition on which it is based on which its success is bound to depend. This supposition is that the aims for which each specialist technique is devised should be understood, not merely by the specialists concerned but by all who wish to make use of their work. This, I submit, is not too much to ask. It does not require that we should immerse ourselves in a study of the technical details of any scientific method ; of the merits of rival methods in the same field ; or of the technical criteria which, in most fields, are gradually developed to test these merits. We are concerned that when a group of specialists claim to have solved, or to have made a step towards solving, one problem, they shall not be taken by their scientific collaborators to have done something else, perhaps of a totally different kind, or perhaps of the same kind, though beyond the reach of their methods or material.

The danger of such confusion is, of course, greatest with those techniques which have been least widely studied, as is the case with statistical methods. For some reason, which the author cannot attempt to explain, even simple statistical methods are not widely taught at the present time, although those needed are certainly not more difficult than much of the same kind which is included in the arithmetic books, and is therefore a part of general knowledge. Moreover, the concepts which these simple methods are fitted to clarify pervade everyday speech, and appear universally in all the observational sciences. We need only recall such phrases as " highly exceptional," " relatively constant," " increases the probability," " adds to our information," " chance fluctuations," and so forth. However, the fact is, whether it be necessary or accidental, that the majority of anthropologists, as of biologists, feel so unfamiliar with statistical reasoning as to accept, in some cases, alleged statistical conclusions with something akin to credulous awe, or in others to reject them with indignation as introducing unnecessary confusion into otherwise plain issues.

 Journal of the Royal Anthropological Institute, 66: 57-63, (1936).

In understanding the Coefficient of Racial Likeness the first thing to be noted is that it is a test of significance. This is a technical term, standing for an idea very prevalent in experimental science, which no one need fail to understand, for it can be made plain in very simple terms. Let us suppose, for example, that we have measurements of the stature of a hundred Englishmen and a hundred Frenchmen. It may be that the first group are, on the average, an inch taller than the second, although the two sets of heights will overlap widely. If the two groups have been chosen from their respective populations in such a way as not to be random samples of the populations they represent, then an examination of the samples will clearly not enable us to compare these populations; but even if our samples are satisfactory in the manner in which they have been obtained, the further question arises as to whether a difference of the magnitude observed might not have occurred by chance, in samples from populations of the same average height. If the probability of this is considerable, that is, if it would have occurred in fifty, or even ten, per cent. of such trials, the difference between our samples is said to be "insignificant." If its probability of occurrence is small, such as one in a thousand, or one in a hundred, or even one in twenty trials, it will usually be termed "significant," and be regarded as providing substantial evidence of an average difference in stature between the two populations sampled. In the first case the test can never lead us to assert that the two populations are identical, even in stature. We can only say that the evidence provided by the data is insufficient to justify the assertion that they are different. In the second case we may be more positive. We know that either our sampling has been exceptionally unfortunate, or that the populations really do differ in the sense indicated by the available data. The chance of our being deceived in the latter conclusion may be very small and, what is more important, may be calculable with accuracy, and without reliance on personal judgment. Consequently, while we require a more stringent test of significance for some conclusions than for others, no one doubts, in practice, that the probability of being led to an erroneous conclusion by the chances of sampling only, can, by repetition or enlargement of the sample, be made so small that the reality of the difference must be regarded as convincingly demonstrated.

It may be asked how we can speak of "how often" a certain average will be recorded in a thousand trials when, in fact, we have only one sample to base our knowledge on. If, indeed, we had a thousand samples of Englishmen, all of the same number, we could, of course, see in how many of them, if in any, the observed average stature was as low as in the French sample. We could do the same *mutatis mutandis* if we had a thousand similar samples of Frenchmen. But, in fact, we have only supposed ourselves to possess one sample from each nation. The point is really one which deserves attention, and the failure to make it clear is certainly responsible for a great part of the misapplication and consequent mistrust from which statistical reasoning has suffered. The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our two hundred

In understanding the Coefficient of Racial Likeness the first thing to be noted is that it is a test of significance. This is a technical term, standing for an idea very prevalent in experimental science, which no one need fail to understand, for it can be made plain in very simple terms. Let us suppose, for example, that we have measurements of the stature of a hundred Englishmen and a hundred Frenchmen. It may be that the first group are, on the average, an inch taller than the second, although the two sets of heights will overlap widely. If the two groups have been chosen from their respective populations in such a way as not to be random samples of the populations they represent, then an examination of the samples will clearly not enable us to compare these populations ; but even if our samples are satisfactory in the manner in which they have been obtained, the further question arises as to whether a difference of the magnitude observed might not have occurred by chance, in samples from populations of the same average height. If the probability of this is considerable, that is, if it would have occurred in fifty, or even ten, per cent. of such trials, the difference between our samples is said to be " insignificant." If its probability of occurrence is small, such as one in a thousand, or one in a hundred, or even one in twenty trials, it will usually be termed " significant," and be regarded as providing substantial evidence of an average difference in stature between the two populations sampled. In the first case the test can never lead us to assert that the two populations are identical, even in stature. We can only say that the evidence provided by the data is insufficient to justify the assertion that they are different. In the second case we may be more positive. We know that either our sampling has been exceptionally unfortunate, or that the populations really do differ in the sense indicated by the available data. The chance of our being deceived in the latter conclusion may be very small and, what is more important, may be calculable with accuracy, and without reliance on personal judgment. Consequently, while we require a more stringent test of significance for some conclusions than for others, no one doubts, in practice, that the probability of being led to an erroneous conclusion by the chances of sampling only, can, by repetition or enlargement of the sample, be made so small that the reality of the difference must be regarded as convincingly demonstrated.

It may be asked how we can speak of " how often " a certain average will be recorded in a thousand trials when, in fact, we have only one sample to base our knowledge on. If, indeed, we had a thousand samples of Englishmen, all of the same number, we could, of course, see in how many of them, if in any, the observed average stature was as low as in the French sample. We could do the same *mutatis mutandis* if we had a thousand similar samples of Frenchmen. But, in fact, we have only supposed ourselves to possess one sample from each nation. The point is really one which deserves attention, and the failure to make it clear is certainly responsible for a great part of the misapplication and consequent mistrust from which statistical reasoning has suffered. The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our two hundred

actual measurements were written on cards, shuffled without regard to nationality, and divided at random into two new groups of a hundred each. This division could be done in an enormous number of ways, but though the number is enormous it is a finite and a calculable number. We may suppose that for each of these ways the difference between the two average statures is calculated. Sometimes it will be less than an inch, sometimes greater. If it is very seldom greater than an inch, in only one hundredth, for example, of the ways in which the sub-division can possibly be made, the statistician will have been right in saying that the samples differed significantly. For if, in fact, the two populations were homogeneous, there would be nothing to distinguish the particular subdivision in which the Frenchmen are separated from the Englishmen from among the aggregate of the other possible separations which might have been made. Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method.

It will be seen that the test of significance does no more, and attempts no more, than to answer the straightforward question, "Could these samples have been drawn at random from the same population?" It calculates a probability. If the probability is very small the answer is "No." If it is not so small as to reach the level of significance required, the answer is "Yes, they could." The answer never is "Yes, they must have been."

It must be stressed that the test of significance calculates a probability; it does not calculate a racial difference. The best estimate of racial difference in stature supplied by the samples, if we are content that it should be called a measure of *racial* difference, is the actual difference observed of one inch. This was the value from which we started. The test of significance does not alter the value, but is a way of examining its reliability. Had we obtained a difference of one inch between samples of only ten statures each, the test would have told us that the difference was insignificant, or, in other words, that samples of ten from the same population would rather often differ by as much. Based on samples of a hundred each the test would probably have reached a level of significance which would satisfy the experimental biologist in the ordinary course of his business. If much larger samples were available they would suffice to demonstrate the statistical significance of much smaller differences. So that while, with meagre data, a difference of a whole inch might be quite insignificant, with more abundant material a difference of only a quarter of an inch or less might be demonstrated beyond reasonable question.

The attempt to discriminate races by measurements of the skull will naturally sometimes require tests having the same purpose, though of a more complex character. The skull may or may not be intrinsically a part of the body particularly well suited for the purpose of distinguishing races, but it has some advantages; it is exceedingly durable, and affords the opportunity of making numerous measurements, with rather high precision. Craniometry attempts to exploit these advantages for ethnographic purposes;

and clearly little progress could be made towards this end unless it were possible to answer the preliminary question whether a given skull or group of skulls could or could not be regarded as a random sample from the same population as some other assemblage of skulls previously measured. If it could be so regarded then measurements of the skull have afforded no means of making a racial discrimination. If it could not, then there is an observable difference of some sort, which may or may not be indicative of a difference in race. The test of significance is of a more complex character than that required for the difference in stature, discussed above, because there are many measurements available, and measurements of the same skull are not generally independent; but its aim and its limitations are the same. The purpose of the Coefficient of Racial Likeness is to afford a means of testing whether two groups of skulls, on each of which numerous measurements have been made, could or could not have been drawn at random from the same bulk. The same question could be answered, often quite decisively, if in any one of the measurements made the two series differ significantly. The advantage of making a combined test lies in the possibility that series which do not differ significantly in any one measurement may yet show differences which, in the aggregate, cannot reasonably be ascribed to chance. The possibility of making a combined test is thus rather a resource useful on special occasions than an ordinary necessity.

As there is some inconsistency in the way in which the term coefficient of racial likeness has been employed, it may be convenient to quote a form which has been widely used.

$$\frac{1}{m}\,\mathrm{S}\left\{\frac{n_s\,n_{s'}}{n_s+n_{s'}}\left(\frac{\mathrm{M}_s-\mathrm{M}_{s'}}{\sigma_s}\right)^2\right\}-1\pm\frac{\cdot 67449}{\sqrt{2\,m}}$$

Here m is the number of measurements used in the test, and S stands for summation of the values in the following bracket for each measurement; n_s and $n_{s'}$ are the numbers of skulls in the two series for which any particular measurement is available, M_s and $\mathrm{M}_{s'}$ the corresponding means, and σ_s is an estimate of the standard deviation within the series which is sometimes obtained from extraneous material, but which in a test of significance should certainly be obtained from the actual measurements of the two series to be compared. The last term is the probable error, to which a coefficient calculated from m really independent measurements would be liable. While the purpose of these calculations is still so questionable, it would be useless to criticise this and similar formulæ in detail. The principal causes of current misunderstanding seem to be the two discussed below.

As has been explained above, it is the function of a test of significance to measure a probability, and not to afford an estimate of a metrical difference. It is, therefore, somewhat unfortunate that the name assigned to the Coefficient of Racial Likeness does suggest, to many who first hear of it, that it affords a measure of the differences, or inversely of the likenesses, between different races. This, of course, it does not attempt

to do,[1] nor is any special statistical device needed for doing this ; for the differences between the averages of each particular measurement in the available samples afford the most direct estimates which we can have of such differences as may exist between the populations sampled.

It should be noted also that, once the level of significance is reached, larger values of the quantity used in testing significance do not alter our conclusions ; they only reiterate the assertion that the populations sampled are in some respects different. They do not show them to be more greatly different. In other words, if we find a value which is not greatly in excess of the limit chosen for significance, we may conclude that the populations really differ, and that the sizes of their samples have been sufficient to prove the existence of this difference. If our value greatly exceeds the limit, it only shows that our samples are larger than would be needed for this purpose. High values of the coefficient of racial likeness do not demonstrate that the races showing them differ more in their cranial measurements than races showing lower, though significant, values.

In the case of the Coefficient of Racial Likeness a second consideration should also prevent any attempt to interpret it as a measure of racial difference, for this coefficient is liable to be large if the series of skulls being compared differ considerably in any measurable feature, whatever that feature may be. In taking it to be a measure of likeness we should be liable to make the assertion that one pair of samples were more like each other than another pair, when in reality they were less like in some features, though perhaps more in others, and for any subsequent utilization of the facts ascertained by measurement it is essential to know in just what respects any two series differ. In any particular application, in fact, we may confidently anticipate that some differences will be unimportant, though others may be genuinely indicative of a racial distinction. Moreover, all skulls have not been measured, and may not be measurable, in respect of exactly the same features. In testing *significance* it is proper to take account of just such measurements as we have, and to examine whether these show any statistically significant differences. But, if we were to mistake the coefficient for an inverse measure of general resemblance, we should be led to suppose that two series of skulls were closely alike on no better ground than that the measurements we possessed of them were few or unimportant.

The first criticism, then, is that the name assigned to the Coefficient of Racial Likeness has led to some misunderstanding of its function and possible use. The second is that as

[1] A quotation from one of the earliest papers in which the Coefficient of Racial Likeness was used will emphasise the limitations I have stressed, and will perhaps throw some light on the origin of later confusions :—

> It is not a true measure of absolute divergence, and must not for a moment be considered as such, but nevertheless we shall speak of it, for convenience, as if it were an absolute measure of racial affinity. When it is said that a low coefficient between two races A and B indicates a closer relationship than a high coefficient between, say, A and C, what is meant always is that it is more probable that A and B are random samples from the same population than that A and C are (pp. 206-207).
>
> —G. M. Morant (1923) *Biom.* xiv. 193-264.

a test of significance it is not a very reliable one. The development of accurate tests of significance must be regarded as one of the more recent developments of statistical technique, and has been rendered possible only by the gradual spread among statisticians of the realisation that testing significance is an entirely different process from making a statistical estimate. During recent years, however, as this has been more and more clearly realised, it has been found possible to make the tests required with precision and rapidly in all cases that ordinarily arise. That of multiple measurements on skulls presents in this respect no special difficulty. The Coefficient of Racial Likeness is defective in that it takes no account of the correlation, or covariation, of different measurements of the same skull, but treats them as though they were statistically independent. The effect of this is to cause very high or very low values of the coefficient to occur more frequently by chance than they should. This effect increases rapidly, both for statistical and for anatomical reasons, as the number of different measurements used is increased, and may, perhaps, account for some of the great contrasts in the value of the coefficient which have been found. In any case in which the coefficient is thought to yield indications of ethnographic importance, which are not apparent by a direct examination of the averages, it is at least advisable to see whether its value has been greatly affected by the mutual correlation of the different measurements.

It will have occurred to the reader who has followed this article so far that the science of craniometry must be in a very primitive condition, if it is still concerned with clarifying its fundamental notions at the stage we have been discussing. It seems, indeed, undoubtedly true that the theoretical concepts developed in the subject have lagged far behind the mass of observational material which has been accumulated. This may be partly due to the sheer magnitude of the programme which the energy of its founders sketched out, partly to an intuitive confidence, widely held in other fields, though everywhere difficult to justify, that, by amassing sufficient statistical material, all difficulties may ultimately be overcome. Partly, again, to an unconscious minimising of these difficulties. For the establishment of statistically reliable differences between series of skulls from different parts of the world, and from different periods, and the further task of evaluating with precision the magnitude of the measurable differences, does not in itself bring us appreciably nearer the stage of recognising which, if any, among our measurements are of the greatest and which are of the least value as indicators of racial affinity. If, indeed, we knew, of each of these measurements, whether it is much or little affected by purely environmental circumstances; and again, whether it has been often and rapidly, or but seldom and that slowly, modified, without racial intermixture, by the selective influences to which human populations are exposed, it might be that some of them, or more probably some particular aspects of the aggregate of measurements, might prove to be of taxonomic value; and to afford genuine quantitative indications of the extent to which primary race stocks, at present largely hypothetical, had been mingled in any particular and observable population. But these necessary and preliminary

enquiries seem largely to have been ignored, for a reason which is not far to seek, namely, the concentration of attention on skeletal remains to the comparative neglect of measurements of living populations.

The durability of skulls has led to their being collected and stored in large numbers ; and they possess a real, though slight, advantage over living heads in the accuracy with which they can be measured. This advantage, though it appeals greatly to the precise worker, is very unimportant, since the variations produced by fleshy tissues are small compared with the metrical differences between individuals, with the consequence that the average of any measurement taken on the living, from a sample of fifty or a hundred, has practically the same precision as that of the corresponding measurement of the skulls. It is on the precision of such averages, and not of the individual measurements, that the possibility of detecting significant differences depends. It may be said also that more measurements are possible with a skull than can be taken on a living head ; but while the truth of this may be granted, it should be pointed out that no inaccessible measurements are known to afford racial discriminations which are not also revealed by external measurements.

On the other hand the living material has some advantages. The sex is known independently of the measurements. Blood relationships are known ; as are nationality, language, religion and social status, all of which greatly affect intermarriage. The heredity of the head measurements has been neglected, curiously when we consider the stress that has been laid on their racial importance. Above all, the student of living measurements can choose his material and be sure of getting enough of it. He is not dependent upon such an accident as that the cemetery of a community, perhaps not well representative of the racial type of its neighbours, has happened to be excavated with unusual thoroughness.

Broadly speaking, and by analogy with the progress of other sciences, it may be suggested that the fundamental problems needed for the ethnographic interpretation of cranial remains must be advanced many steps further than the present state of knowledge before they can contribute appreciably to our knowledge of racial history. And that the material for throwing light on these fundamental problems is only to be found in the examination of existing living populations.