From the Journal and Proceedings, Asiatic Society of Bengal (New Series), Vol. XXVI, 1930, No. 4.

ARTICLE No. 36.

On Tests and Measures of Group Divergence.

Part I: Theoretical Formulæ

By PRASANTA CHANDRA MAHALANOBIS

CONTENTS

I. INTRODUCTION

1. In many statistical investigations two important questions arise in dealing with two or more "samples". Let S and S' be two given samples. Then either of two things may have happened:—

(A) both S and S' were drawn from the same group G,[1]

or

(B) the samples S and S' were drawn from two different groups G and G'[2].

---

[1] I have used the word "group" in the present paper in very nearly the same sense as the word "population" is used in statistical literature. A "group" will denote any collection of individuals or entities; the individuals (constituting the group) may be distinguished from one another, but all of them possess certain common characteristics, by virtue of which common characteristics they are supposed to belong to the same "group". I have reserved the word "population" for use in a slightly more general sense; so that when necessary we shall be able to speak of the existence of "groups" within a population. A "sample" is an aggregate of measurements, in one or more specified characters, of a finite number of individuals belonging to the same group (or population). It is throughout assumed in the present paper that all samples are "random" samples, *i.e.*, the individuals constituting the sample are not selected in any way, and are drawn in a random manner from the group or population concerned.

[2] A third alternative is that S and S' were both drawn from the same group (or population) but either or both of them were selected samples. This hypothesis is however excluded by our assumption that all samples are random samples. (See footnote 1.)

Any criterion which will distinguish between (A) and (B) *i.e.*, determine whether the two given samples are drawn from the same or from two different groups (or populations) may be called a *test* of group divergence.

In case the two samples S and S' are considered to be drawn from two different groups G and G', it is obvious that G' may be any one of an infinite number of groups differing only slightly or very greatly from G. Any coefficient which would furnish information regarding the actual amount of the divergence subsisting between G and G' may be called a *measure* of group divergence. The distinction between a "test" and a "measure" of group divergence is fundamental; a test merely tells us whether the two groups (from which the two given samples are drawn) are different or not, while a "measure" gives us a quantitative estimate of the magnitude of the difference (if any) between the two groups.

2. *Notation.* In dealing with more than one group, it is necessary to distinguish carefully between different type of means and standard deviations.

Let x_{pqt} represent a single measurement of the t^{th} individual in the q^{th} sample for the p^{th} character, and let n_{pq} be the total number of individuals in the q^{th} sample for the p^{th} character, N_p the total number of samples available for the p^{th} character, and P the total number of characters for which measurements were taken.

The total number of individuals for whom measurements of the p^{th} character are available will be given by

$$n_p = S_q[(n_{pq})] \quad .. \quad .. \quad .. \quad .. \quad (2{\cdot}1)$$

where S_q denotes a summation for all samples, *i.e.*, for all values of q (from $q=1$, to $q=N_p$). When n_{pq} is constant for all samples (*i.e.*, for all values of q),

$$n_p = N_p \cdot n_{pq}. \quad .. \quad .. \quad .. \quad .. \quad (2{\cdot}2).$$

The *intra-class mean* (m_{pq}) and the *intra-class variance* (σ^2_{pq}) for the q^{th} sample in the p^{th} character are defined by:—

$$n_{pq} \cdot m_{pq} = S_t [(x_{pqt})] \quad .. \quad .. \quad .. \quad .. \quad (2{\cdot}3)$$

$$n_{pq} \cdot \sigma^2_{pq} = S_t [(x_{pqt} - m_{pq})^2] \quad .. \quad .. \quad (2{\cdot}4)$$

where S_t is a summation for all individuals within the given sample, *i.e.*, for all values of t.

For the p^{th} character there will be N_p such means and N_p such variances, one pair for each of the N_p different samples.

The *inter-class means* (M_p) and the *inter-class variance* (s_p^2) for the p^{th} character are defined by:—

$$N_p \cdot M_p = S_q [(m_{pq})] \quad .. \quad .. \quad .. \quad (2{\cdot}5)$$

$$N_p \cdot s_p^2 = S_q [(m_{pq} - M_p)^2]. \quad .. \quad .. \quad (2{\cdot}6).$$

If all the individuals are pooled together for any particular character we shall get another set of means (m_p) and variance (Σ_p^2) defined by :—

$$n_p \cdot m_p = S_q [(n_{pq} \cdot m_{pq})] = S_q S_t [(x_{pqt})] \quad .. \quad (2{\cdot}7)$$

$$n_p \cdot \Sigma_p^2 = S_q S_t [(x_{pqt} - m_p)^2]. \quad .. \quad .. \quad .. \quad (2{\cdot}8).$$

Following a suggestion of Prof. Karl Pearson such means and variances may be called the "familial" means and variances.

Besides the above we may also define an *average intra-class* variance by :—

$$N_p \cdot \bar{\sigma}_p^2 = S_q [(\sigma^2_{pq})]. \quad .. \quad .. \quad .. \quad .. \quad (2{\cdot}9).$$

The mean, variance, etc., of the group (or population) from which a sample is drawn may be written as $\bar{m}_{pq}$, $\bar{\sigma}^2_{pq}$, etc. When there is no chance of confusion, for example, for only two samples in any assigned character, we may drop the subscripts and write m and m' for the intra-class means, σ^2 and σ'^2 for the intra-class variances, n and n' for the size of the two samples, and $\bar{m}$ and $\bar{m'}$ for the two corresponding group-means.

I shall write dm, dm', etc., everywhere for statistical differences (*i.e.*, deviations of individuals' values from corresponding mean values).

II. Tests of Divergence

3. *Single character* ($P=1$). When n, n' are both large, (say greater than 25), it is often possible to use the normal (Gauss-Laplacian) distribution of deviations. It will be only necessary to calculate the statistics

$$x = (m - m') \sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}} \quad .. \quad .. \quad .. \quad (3{\cdot}1)$$

and using a standard table of the probability integral (10, pp. 2–8)[1] calculate the probability of occurrence of a deviation equal to or greater than "x."

But if the size of the group is small, (*e.g.*, when n and n' are less than 25), the method given by R. A. Fisher (2, p. 107) may be used with advantage, especially when there are reasons for believing that there is no significant difference in the variability of the two samples. Two statistics are calculated, one the "pooled" variance given by

$$s^2 = \frac{(n-1)\sigma^2 + (n'-1)\sigma'^2}{(n-1) + (n'-1)} \quad .. \quad .. \quad .. \quad (3{\cdot}2)$$

---

[1] The number within brackets refers to the list given at the end.

and the other, the deviation

$$t=\left(\frac{m-m'}{s}\right)\sqrt{\frac{n.n'}{n+n'}} \quad .. \quad .. \quad .. \quad (3\cdot3)$$

and the probability of occurrence of deviations as great or greater than "t" is obtained from tables given by Fisher (2, p. 139).

In other cases certain tests recently developed by J. Neyman and E. S. Pearson (8) may be used. The most convenient test in practice would probably be the $P\lambda$ test for which necessary tables have been supplied by the authors.

A general treatment of the problem is also possible which takes into consideration the nature of the frequency distribution as a whole, Pearson (11) has shown that if f_p and f'_p are the frequencies in corresponding cells for two samples (both of which are supposed to be random samples drawn from the same general population), of sizes n and n' respectively, then on the assumption that there is no correlation of deviations in frequencies between the first and the second group, the statistics

$$\chi^2=S_p\left[\frac{n.n'\left(\frac{f_p}{n}-\frac{f_p'}{n'}\right)^2}{(f_p+f_p')}\right] \quad .. \quad .. \quad .. \quad (3\cdot4)$$

(where S_p denotes a summation for all p cells), may be used for obtaining the probability of occurrence of the given system of differences from standard tables (10, Table XII, p. 26).

4. *Multiple characters.* When the number of characters is more than one, the Pearsonian Coefficient of Racial Likeness furnishes the standard test of divergence.

$$C^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\left(\frac{\sigma_{pq}^2}{n_{pq}}+\frac{\sigma_{pq}'^2}{n_{pq}'}\right)}\right]-1 \quad .. \quad .. \quad .. \quad (4\cdot0)$$

Where P is the total number of characters for which the summation is taken. This coefficient was first used by Miss M. L. Tildesley (14, p. 247) in 1921, and later on extensively by Dr. G. H. Morant (7) and others. Prof. Pearson (12) gave a full theoretical discussion in 1926.

If the two samples are both random samples drawn from the same general population, then the theoretical value of C^2 is given by

$$(C^2)_0=0\pm\cdot67449\sqrt{\frac{2}{P}} \quad .. \quad .. \quad .. \quad .. \quad (4\cdot1).$$

If C^2 differs significantly from zero then the two samples cannot be considered to be random samples drawn from the same population.

If we use a reliable constant value[1] of the intra-group variance $\overline{\sigma_p}^2$ in the place of the observed values σ_{pq}^2 and $\sigma_{pq'}^2$, C^2 may be written as

$$C^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq'})^2}{\overline{\sigma_p}^2\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq'}}\right)}\right]-1 \quad .. \quad .. \quad (4{\cdot}01)$$

or

$$C^2=\left(\frac{n.n'}{n+n'}\right)\cdot\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq'})^2}{\overline{\sigma_p}^2}\right]-1 \quad .. \quad (4{\cdot}02)$$

when the size of the samples is constant for all characters.

5. It will be noticed that C^2 is an adequate test of divergence only so far as group-means are concerned. It is obvious that two groups may agree in their means and yet be divergent in other characteristics such as variance, skewness or kurtosis. Separate tests of divergence for such other characteristics are therefore necessary, and may be easily constructed.

For example for testing divergence in variability we may use the following coefficient

$$E^2=\frac{2}{P}S_p\left[\left(\frac{n_{pq}\cdot n_{pq'}}{n_{pq}+n_{pq'}}\right)\cdot\frac{(\sigma_{pq}-\sigma_{pq'})^2}{\overline{\sigma_p}^2}\right]. \quad .. \quad (5{\cdot}1).$$

Proceeding in the same way we can test the divergence in skewness or kurtosis with the help of the following coefficients G^2 and H^2 respectively.

$$G^2=\frac{1}{P}\,S_p\left[\frac{(sk-sk')^2}{\Sigma_{sk}^2+\Sigma_{sk'}^2}\right]-1 \quad .. \quad .. \quad .. \quad (5{\cdot}2)$$

$$H^2=\frac{1}{P}\,S_p\left[\frac{(\beta_2-\beta_2')^2}{\Sigma\beta_2^2+\Sigma\beta_2'^2}\right]-1 \quad .. \quad .. \quad .. \quad (5{\cdot}3)$$

where sk, sk' the two skewness, with their variances Σ_{sk}^2, $\Sigma_{sk'}^2$, and β_2, β_2' with the corresponding variances $\Sigma\beta_2^2$, $\Sigma\beta_2'^2$ can be easily obtained from equations and tables given by Pearson (10, Tables XXXV–XLVI, pp. 66–87).

When the two groups are random samples drawn from the same population, the mean values of E^2, G^2, and H^2 will be each equal to

$$(E^2,\ G^2,\ H^2)=0\pm{\cdot}67449\sqrt{\frac{2}{P}} \quad .. \quad .. \quad (5{\cdot}4).$$

---

[1] An estimate based on a long series of measurements may be used for this purpose; or where the intra-class variance of a fairly large number of samples are known, an average value of the intra-class variance as defined by (2·9) can be easily calculated.

It is only when all these coefficients (in addition to C^2) are sensibly zero shall we be justified in asserting that there is no divergence between the two groups (up to the order of the 4th moment).

In actual practice it will be often difficult to use G^2 or H^2 as the estimates of the variances will usually be unreliable owing to the smallness of the size of the samples. It should, however, be possible to use E^2 in many cases. Numerical examples will be found in Part II of the present paper.

6. A more general discussion based on the method of paragraph 3 is theoretically possible. The two groups to be compared may be subdivided into a large number of cells in a P-dimensional manifold, and the frequencies in each cell may be used for the calculation of χ^2 defined by equation (3·4).

If c_1, c_2, c_3,.... c_p... are the number of sub-classes into which the 1st, the 2nd, the 3rd,... the p*th* character is split up, then the total number of elementary cells will be given by $c_1 . c_2 . c_3 .... c_p = c$.

If we use very broad categories, say only 4 divisions for each character, then $c_1 = c_2 = c_3 = ... c_p = 4$.

Thus $c = 4^P$, the total number of cells, will become very large even for small values of P, and hence it will become impracticable to use the present method in most cases.

III. Measures of Divergence

7. The entity we have been calling "the amount of divergence between two groups" is a derived quantity, and is not given directly. A certain amount of choice exists in its precise formulation, and its exact significance will depend upon and will be determined by the particular mathematical formula by which we choose to define it.

If each sample is represented by a point in a P-dimensional manifold determined by P values of the means of P characters, what we obviously require is a suitable expression for an entity which may be called the generalised (P-dimensional) distance between any pair of such P-fold points. If all the characters were directly comparable, we could use the ordinary quadratic expression $S_p [(m_{pq} - m_{pq'})^2]$. But we are confronted with the difficulty that all the characters are not directly comparable.

The crux of the whole problem lies, therefore, in transforming the raw observed differences $(m_{pq} - m_{pq'})$ in such a way that they may all become directly comparable with one another. It is clear therefore that we must introduce suitable multipliers or "weights," so that a difference, say $(m_{1,q} - m_{1',q'})$ in one character will, in some defined sense, match or be equivalent to a corresponding $(m_{2,q} - m_{2',q'})$ in a second character.

Introducing k_p as a suitably selected multiplier, we obtain the general form for a measure of divergence in means:

$$U^2 = f\left(S_p\left[\frac{(m_{pq} - m_{pq}')^2}{k_p^2}\right]\right) \quad \ldots \quad \ldots \quad \ldots \quad (7{\cdot}0).$$

It will be convenient to choose k_p in such a way that it may satisfy the following conditions:—

(i) U^2 should be a pure number. This requires that k_p should have the same dimensions as m_{pq} or m_{pq}'.

(ii) U^2 should vanish when the samples are both random samples drawn from the same general population.

(iii) U^2 should be constant (within the limits of errors of random sampling) for two samples drawn from the same two differing groups or populations.

(iv) U^2 should increase (or decrease) as the system of differences $(m_{pq} - m_{pq}')$ increases (or decreases). For example, for 3 groups G_1, G_2, and G_3, if it actually happens in practice that for all characters the differences in means between the 1st and the 2nd group, *i.e.*, the quantities $(m_{p,1} - m_{p,2})$ are less than $(m_{p,1} - m_{p,3})$ the corresponding differences between the 1st and the 3rd group, then U_{12}^2 the divergence between G_1 and G_2 should be less than U_{13}^2 the divergence between G_1 and G_3. This condition suggests that k_p should be kept invariable (for each character) for the same series of comparisons.

In choosing k_p we must be guided by empirical considerations; recourse to a method of trial and error is, therefore, inevitable. To this extent the choice of k_p is arbitrary, *i.e.*, we are free (in fact we are obliged) to try different values of k_p, and see what kind of results emerge from each value of k_p so chosen. The ultimate choice, however, will be determined or (limited) by the actual facts of nature. For we must finally adopt that particular value of which will yield in practice a system of description possessing the greatest coherence, range, significance, and simplicity.

In the case of anthropology it is conceivable that genetic analysis may develop far enough in future to be able to furnish us with a reliable set of values of k_p for different characters. But for the present, restricting ourselves to purely statistical considerations, the choice of k_p would appear to lie among two groups of constants.

(*a*) We may choose one or other of the different measures of variation:—

(i) the intra-class standard deviation (σ_p);
(ii) the inter-class standard deviation (s_p); or,
(iii) the familial standard deviation (Σ_p).

All the above quantities have the advantage that they can be determined with greater "efficiency" in the sense defined by

Fisher (3) than the absolute range,[1] the mean deviation, or one of the percentile differences. I have, therefore, confined my discussion to the three standard deviations.

(*b*) In the alternative we can use the inter-class mean M_p or the familial mean m_p. It is obvious that so far as anthropometry is concerned both would give practically the same results, as the difference between the two will in most cases be negligible in comparison with the magnitude of either.

8. Similar considerations will apply to the case of other group constants such as the variance, the skewness, or the kurtosis. In fact we can generalise equation (7·0), and write for *any measure of divergence:*

$$U^2=f\left(S_p\left[\frac{(x_{pq}-x_{pq}')^2}{K_p^2}\right]\right) \quad .. \quad .. \quad .. \quad (8\cdot0)$$

where x_{pq}, x_{pq}' are corresponding values of the same statistical entity for *qth* and *q'th* samples respectively, and K_p^2 is a suitably chosen multiplier which does not involve either x_{pq} or x_{pq}'.

9. Before proceeding further it will be useful to obtain a few statistical formulæ connected with equation (8).

Let x, x' be the observed values of any particular statistics for two samples of size n and n' respectively. Let $\bar{x}$, $\bar{x}'$ be the corresponding values of the same statistics for the two respective groups (or populations) from which the two samples are drawn.

We shall assume[2] that these "true" or "mean" values may be reached by taking the average of an indefinitely large number of samples. .. .. (A·1).

Let us write

$$z=(x-x'). \quad .. \quad .. \quad .. \quad .. \quad (9\cdot0).$$

If $\bar{z}$ is the "true" or "mean" values of z, (as defined above), and dz, dx, dx' are statical deviations from the corresponding mean values, then we may write:

$$z=\bar{z}\left(1+\frac{dz}{\bar{z}}\right)=\bar{x}\left[\left(1+\frac{dx}{\bar{x}}\right)-\bar{x}'\left(1+\frac{dx'}{\bar{x}'}\right)\right] \quad .. \quad (9\cdot01).$$

Squaring we get

$$\bar{z}^2\left(1+2\frac{dz}{\bar{z}}+\frac{dz^2}{\bar{z}^2}\right)=\left[\bar{x}\left(1+\frac{dx}{\bar{x}}\right)-\bar{x}'\left(1+\frac{dx'}{\bar{x}'}\right)\right]^2 \quad .. (9\cdot02).$$

We can easily find the mean value ($\bar{z}$) and (σ_z^2) the variance of z, if we make certain simple assumptions:—

---

[1] The "range" (being the difference in character between the largest and the smallest individuals of a sample) is not suitable owing to the fact that its value depends on the size of the sample. [See Refs. 9 and 16.]

[2] Assumptions are clearly indicated by separate serial numbers: (A·1), (A·2) etc.

(A·2). The distribution of (dx) and (dx') are both normal. so that summing and taking the average for a very large number of samples, and using $\{\ \}$ brackets for writing such average values we have

$$\left.\begin{aligned}&\{dx\}=\{dx'\}=\{dx^3\}=\{dx^5\}=\ldots.\{dx^{2r+1}\}=0\\&\{dx^2\}=\sigma_x^2,\{dx^4\}=3\sigma_x^4,\{dx^6\}=15\sigma_x^6,\{dx^8\}=105\sigma_x^8,\text{ etc.}\end{aligned}\right\}(A\cdot2)$$

with a similar set of expressions for dx' where σ_x^2, $\sigma_x'^2$ are the variance of x and x' respectively.

(A·3). The deviations (dx) and (dx') are statistically independent, so that the product terms involving odd-powers will vanish, and product terms involving even-powers may be summed independently.

$$\left.\begin{aligned}&\{dx\,.\,dx'\}=\{dx^2\,.\,dx'\}=\{dx\,.\,dx'^2\}=\ldots.=0\\&\{dx^2.\,dx'^2\}=\sigma_x^2\,.\,\sigma_x'^2,\ \{dx^4\,.\,dx'^2\}=3\sigma_x^4\,.\,\sigma_x'^2,\\&\qquad\qquad\{dx^2\,.\,dx'^4\}=3\sigma_x^2\,.\,\sigma_x'^4,\text{ etc.}\end{aligned}\right\}\ldots\ (A\cdot3)$$

Summing and taking the average for a very large number of samples for equations (9·01) and (9·02) we easily find

$$\bar{z}=(\bar{x}-\bar{x}') \quad \ldots\ \ldots\ \ldots\ \ldots \quad (9\cdot1)$$

$$\sigma_z^2=\{dz^2\}=\sigma_x^2+\sigma_x'^2 \quad \ldots\ \ldots \quad (9\cdot2).$$

Taking the cube, and higher powers of equation (9·01) it can be shown in the same way that all the odd moments vanish :—

$$\{dz^3\}=\{dz^5\}=\{dz^7\}=\ldots.0$$

and the even moments are the same as for a normal distribution :—

$$\{dz^4\}=3\sigma_z^4,\ \{dz^6\}=15\sigma_z^6,\ \{dz^8\}=105\sigma_z^8,\ \{dz^{10}\}=945\sigma_z^{10},\text{ etc.}$$

10. Let y be any other statistical quantity whose "true" or mean value is $\bar{y}$. We define a new quantity "a" by

$$a=\frac{(x-x')^2}{y^2}=\frac{z^2}{y^2}\ldots\ \ldots\ \ldots\ \ldots \quad (10\cdot0).$$

If $\bar{a}$ is the mean value of a, and da, dz, dy are statistical deviation from corresponding mean values, we may write

$$a=\bar{a}\left(1+\frac{da}{\bar{a}}\right)=\frac{\bar{z}^2\left(1+\frac{dz}{\bar{z}}\right)^2}{\bar{y}^2\left(1+\frac{dy}{\bar{y}}\right)^2}=a_0\left(1+\frac{dz}{\bar{z}}\right)^2.\left(1+\frac{dy}{\bar{y}}\right)^{-2}\ldots\ (10\cdot1),$$

where

$$a_0\equiv\frac{\bar{z}^2}{\bar{y}^2}=\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}.\ \ldots\ \ldots\ \ldots \quad (10\cdot01).$$

Taking the square, the cube and the 4th power of equation (10·01), we have

$$\bar{a}^2\left(1+2\frac{da}{\bar{a}}+\frac{da^2}{\bar{a}^2}\right) = a_0{}^2.\left(1+\frac{dz}{\bar{z}}\right)^4.\left(1+\frac{dy}{\bar{y}}\right)^4 \quad .. \quad .. \quad (10\cdot2)$$

$$\bar{a}^3\left(1+3\frac{da}{\bar{a}}+3\frac{da^2}{\bar{a}^2}+\frac{da^3}{\bar{a}^3}\right) = a_0{}^3\left(1+\frac{dz}{\bar{z}}\right)^6.\left(1+\frac{dy}{\bar{y}}\right)^6 \quad .. \quad .. \quad (10\cdot3)$$

$$\bar{a}^4\left(1+4\frac{da}{\bar{a}}+6\frac{da^2}{\bar{a}^2}+4\frac{da^3}{\bar{a}^3}+\frac{da^4}{\bar{a}^4}\right) = a_0{}^4\left(1+\frac{dz}{\bar{z}}\right)^8.\left(1+\frac{dy}{\bar{y}}\right)^8 \quad .. \quad .. \quad (10\cdot4)$$

We now make two further assumptions :—

(A·4). The distribution of (y) is normal, so that

$$\left.\begin{aligned}&\{dy\}=\{dy^3\}=\{dy^5\}=\ldots\ldots\ldots\ldots=0\\&\{dy^2\}=\sigma_y{}^2,\ \{dy^4\}=3\sigma_y{}^4,\ \{dy^6\}=15.\sigma_y{}^6,\text{ etc.}\end{aligned}\right\} \quad .. \quad .. \quad (A\cdot4)$$

(A.5). The deviations (dy) and (dz) are statistically independent. This is equivalent to the assumption that (dy) and (dx), as also (dy) and (dx') are statistically independent.

$$\left.\begin{aligned}&\{dz.dy\}=\{dz^2.dy\}=\{dz.dy^2\}=\ldots\ldots=0\\&\{dz^2.dy^2\}=\sigma_z{}^2.\sigma_y{}^2,\ \{dz^4.dy^2\}=3\sigma_z{}^4.\sigma_y{}^2,\text{ etc.}\end{aligned}\right\} \quad .. \quad .. \quad (A\cdot5)$$

We can now expand equations (10·1), (10·2), (10·3), and (10·4) in ascending powers of $\left(\frac{dz}{\bar{z}}\right)^2$ and $\left(\frac{dy}{\bar{y}}\right)^2$, since the odd powers will vanish on taking the average of an indefinitely large number of samples.

We shall write

$$v^2=\frac{\sigma_z{}^2}{\bar{z}^2}=\frac{\sigma_x{}^2+\sigma_{x'}{}^2}{(\bar{x}-\bar{x}^1)^2},\quad w^2=\frac{\sigma_y{}^2}{\bar{y}^2} \quad .. \quad .. \quad .. \quad .. \quad (10\cdot5)$$

and assume that

$$v^2<1,\ w^2<1 \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (A\cdot6)$$

so that we may expand in ascending powers of v^2 and w^2.

The moment coefficients of a may be written as usual :—

$$\{da\}=0,\ \{da^2\}=\mu_2(a),\ \{da^3\}=\mu_3(a),\text{ and }\{da^4\}=\mu_4(a).$$

By straightforward algebra[1] we then obtain the following equations :—

$$\bar{a}=a_0(1+v^2)(1+3w^2+15w^4+105w^6+945w^8+10,395w^{10}) \quad .. \quad .. \quad (10{\cdot}6)$$

$$\mu_2(a)=2a_0{}^2[v^2(2+v^2)(1+12w^2+138w^4+1,740w^6+24,615w^8)+w^2(2+33w^2+480w^4+7,290w^6+120,330w^8)] \quad .. \quad .. \quad (10{\cdot}71)$$

$$=2a_0{}^2[2(v^2+w^2)+(v^4+24v^2w^2+33w^4)+12w^2(v^4+23v^2w^2+40w^4)+6w^4(23v^4+580v^2w^2+1215w^4)+10w^6(1,740v^4+4,923v^2w^2+12,033w^4)] \quad .. \quad .. \quad (10{\cdot}72)$$

$$\mu_3(a)=8a_0{}^3[v^4(3+v^2)+3w^4(3+111w^2+3,030w^4+76,950w^6)+3v^2w^2(4+123w^2+2,833w^4+64,320w^6)+3v^2w^2\{9v^2(3+62w^2+1,232w^4)+3v^4(3+62w^2)\}] \quad .. \quad .. \quad (10{\cdot}81)$$

$$=8a_0{}^3[3(v^4+4v^2w^2+3w^4)+(v^6+81v^4w^2+369v^2w^4+333w^6)+9w^2(3v^6+186v^4w^2+961v^2w^4+1,010w^6)+18w^4(31v^6+1,848v^4w^2+10.720v^2w^4+12,825w^6)] \quad .. \quad .. \quad (10{\cdot}82)$$

$$\mu_4(a)=12a_0{}^4[v^4(4+20v^2+5v^4)+w^4(4+340w^2+16,101w^4+619,560w^6)+2v^2w^2\{2(2+145w^2+6,016w^4+206,556w^6)+v^2(130+4,833w^2+149,904w^4)+8v^4(54+1,665w^2)\}] \quad .. \quad .. \quad (10{\cdot}91)$$

$$=12a_0{}^4[4(v^2+w^2)^2+20(v^6+13v^4w^2+29v^2w^4+17w^6)+(5v^8+864v^6w^2+9,666v^4w^4+2,464v^2w^6+16,101w^8)+24w^2(9v^8+1,110v^6w^2+12,492v^4w^4+34,426v^2w^6+25,815w^8)] \quad .. \quad .. \quad (10{\cdot}92)$$

When y is constant, i.e. when $w_{y}{}^2=0$, (or in any case where the variance of y is negligibly small) we have

$$\bar{a}=\frac{\bar{z}^2}{\bar{y}^2}\cdot(1+v^2)=\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}\cdot\left\{1+\frac{\sigma_x{}^2+\sigma_x{}'^2}{(\bar{x}-\bar{x}^1)^2}\right\}$$

$$=\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}+\frac{\sigma_x{}^2+\sigma_x{}'^2}{\bar{y}^2} \quad .. \quad .. \quad .. \quad (10{\cdot}63)$$

$$\mu_2(a)=2\,\frac{\bar{z}^4}{\bar{y}^4}\cdot v^2(2+v^2)$$

$$=4\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}\cdot\frac{(\sigma_x{}^2+\sigma_x{}'^2)}{\bar{y}^2}+2\,\frac{(\sigma_x{}^2+\sigma_x{}'^2)^2}{\bar{y}^4} \quad .. \quad .. \quad (10{\cdot}73)$$

---

[1] I am indebted to my pupil Mr. Ananda Chandra Ray for verifying some of the algebraic results.

$$\mu_3(a)=8\frac{\bar{z}^6}{\bar{y}^6}\cdot v^4(3+v^2)$$

$$=24\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}\cdot\frac{(\sigma_x^2+\sigma_x'^2)^2}{\bar{y}^2}+8\frac{(\sigma_x^2+\sigma_x'^2)^3}{\bar{y}^6} \quad .. \quad (10{\cdot}83)$$

$$\mu_4(a)=12\frac{\bar{z}^8}{\bar{y}^8}\cdot v^4(4+20v^2+5v^4)$$

$$=48\frac{(\bar{x}-\bar{x}')^4}{\bar{y}^4}\cdot\frac{(\sigma_x^2+\sigma_x'^2)^2}{\bar{y}^4}+360\frac{(\bar{x}-\bar{x}')^2}{\bar{y}^2}\cdot\frac{(\sigma_x^2+\sigma_x'^2)^3}{\bar{y}^6}$$

$$+60\frac{(\sigma_x^2+\sigma_x'^2)^4}{\bar{y}^8} \quad .. \quad .. \quad .. \quad (10{\cdot}93)$$

Also $\beta_1=\mu_3^2/\mu_2^3=8v^2(9-\frac{1}{2}^5\ v^2+\frac{1}{2}^1\ v^4-\frac{1}{2}^5\ v^6)$ .. .. (10·94)

$\beta_2=\mu_4/\mu_2^2=3(1+4v^2-3v^4+2v^6-\frac{5}{1}\cdot v^8)$ .. .. (10·95).

11. We now define

$$b=\frac{1}{P}S_p[a_1+a_2+\ldots]=\frac{1}{P}S_p[a_p] \quad .. \quad .. \quad .. \quad (11{\cdot}0)$$

where $a_1, a_2, a_3, \ldots . a_p$ are each defined by an equation of the type (10·0).

Writing $da_1, da_2, da_3, \ldots . da_p . .$, and db as statistical deviations from the corresponding mean values $\bar{a}_1, \bar{a}_2, \ldots \bar{a}_3$, and $\bar{b}$ respectively, we have

$$\bar{b}\left(1+\frac{db}{\bar{b}}\right)=\frac{1}{P}S_p\left[\bar{a}_p\left(1+\frac{da_p}{\bar{a}_p}\right)\right] \quad .. \quad .. \quad .. \quad (11{\cdot}01).$$

Taking the average value of an indefinitely large number of samples,

$$\bar{b}=\frac{1}{P}S_p[\bar{a}_p] \quad .. \quad .. \quad .. \quad .. \quad (11{\cdot}1).$$

Squaring equation (11·01), we get

$$\bar{b}^2\left(1+2\frac{db}{\bar{b}}+\frac{db^2}{\bar{b}^2}\right)=\frac{1}{P^2}S_p\left[\bar{a}_p^2\left(1+2\frac{da_p}{\bar{a}_p}+\frac{da_p^2}{\bar{a}_p^2}\right)\right]$$

$$+\frac{1}{P^2}S_pS_r\left[2\bar{a}_p\cdot\bar{a}_r\left(1+\frac{da_p}{\bar{a}_p}+\frac{da_r}{\bar{a}_r}+\frac{da_p\cdot da_r}{\bar{a}_p\cdot\bar{a}_r}\right)\right] \quad .. \quad .. \quad (11{\cdot}02)$$

where S_pS_r denotes a summation for all pairs of values of p and r, $(p\neq r)$.

We next assume that a's are statistically independent, i.e., (da_p), (da_r) are statistically independent for all values of p and r, so that

$$(p\neq r), \quad \left.\begin{array}{l}\{da_p\cdot da_r\}=\{da_p^2\cdot da_r\}=\ldots .=0\\ \{da_p^2\cdot da_r^2\}=\mu_2(a_p)\cdot\mu_2(a_r);\ \text{etc.}\end{array}\right\} \quad .. \quad (A{\cdot}7).$$

Taking the average of an indefinitely large number of samples for equation (11.02), and writing $\mu_2(b)$ as the second moment-coefficient of b, we have

$$\bar{b}^2+\mu_2(b)=\frac{1}{P^2}S_p[\bar{a}_p{}^2]+\frac{1}{P^2}S_pS_r[2\bar{a}_p\,.\,\bar{a}_r]+\frac{1}{P^2}S_p[\mu_2(a_p)]\;..(11{\cdot}021).$$

Eliminating $\bar{b}$ with the help of equation (11.1) we obtain

$$\mu_2(b)=\frac{1}{P^2}S_p[\mu_2(a_p)] \quad .. \quad .. \quad .. \quad (11{\cdot}2).$$

Again taking the 3rd power of equation (11.01),

$$\bar{b}^3\left(1+3\frac{db}{\bar{b}}+3\frac{db^2}{\bar{b}^2}+\frac{db^3}{\bar{b}^3}\right)=\frac{1}{P^3}S_p\left[\bar{a}_p{}^3\left(1+3\frac{da_p}{\bar{a}_p}+3\frac{da_p{}^2}{\bar{a}_p{}^2}+\frac{da_p{}^3}{\bar{a}_p{}^3}\right)\right]+\frac{1}{P^3}S_pS_r\left[3\bar{a}_p{}^2\,.\,\bar{a}_r\left(1+2\frac{da_p}{\bar{a}_p}+\frac{da_p{}^2}{\bar{a}_p{}^2}\right)\left(1+\frac{da_r}{\bar{a}_r}\right)\right] \quad .. \quad .. \quad (11{\cdot}03).$$

Writing $\mu_3(b)$ as the 3rd moment-coefficient of b, and taking the average of an indefinitely large number of samples,

$$\bar{b}^3+3\bar{b}\,.\,\mu_2(b)+\mu_3(b)=\frac{1}{P^3}S_p[\bar{a}_p{}^3]+\frac{1}{P^3}S_pS_r[3\bar{a}_p{}^2\,.\,\bar{a}_r]+\frac{1}{P^3}S[3\bar{a}_p\,.\,\mu_2(a_r)]+\frac{1}{P^3}S_pS_r[3\bar{a}_p\,.\,\mu_2(a_r)]+\frac{1}{P^3}S_p[\mu_3(a_p)] \quad .. \quad .. \quad (11{\cdot}031).$$

Using (11.1) and (11.2) we therefore obtain

$$\mu_3(b)=\frac{1}{P^3}S_p[\mu_3(a_p)] \quad .. \quad .. \quad .. \quad (11{\cdot}3).$$

Now taking the 4th power of (11.01),

$$\bar{b}^4\left(1+4\frac{db}{\bar{b}}+6\frac{db^2}{\bar{b}^2}+4\frac{db^3}{\bar{b}^3}+\frac{db^4}{\bar{b}^4}\right)=\frac{1}{P^4}S_p\left[\bar{a}_p{}^4\left(1+4\frac{da_p}{a_p}+6\frac{da_p{}^2}{\bar{a}_p{}^2}+4\frac{da_p{}^3}{\bar{a}_p{}^3}+\frac{da_p{}^4}{\bar{a}_p{}^4}\right)\right]+\frac{1}{P^4}S_pS_r\left[4\bar{a}_p{}^3\,.\,\bar{a}_r\left(1+3\frac{da_p}{\bar{a}_p}+3\frac{da_p{}^2}{\bar{a}_p{}^2}+\frac{da_p{}^3}{\bar{a}_p{}^3}\right)\left(1+\frac{da_r}{\bar{a}_r}\right)\right]+\frac{1}{P^4}S_pS_r\left[6\bar{a}_p{}^2+\bar{a}_r{}^2\left(1+2\frac{da_p}{\bar{a}_p}+\frac{da_p{}^2}{\bar{a}_p{}^2}\right)\left(1+2\frac{da_r}{\bar{a}_r}+\frac{da_r{}^2}{\bar{a}_r{}^2}\right)\right] \quad .. \quad .. \quad (11{\cdot}04).$$

Writing $\mu_4(b)$ as the 4th moment-coefficient of b, taking average values, and eliminating $(\bar{b})$, $\mu_2(b)$, and $\mu_3(b)$ with the help of (11.1), (11.2) and (11.3), we have finally

$$\mu_4(b)=\frac{1}{P^4}S_p[\mu_4(a_p)]+\frac{6}{P^4}S_pS_r[\mu_2(a_p)\cdot\mu_2(a_r)] \quad .. \quad (11{\cdot}4).$$

The above results could of course have been obtained from the more general formulæ given by Professor Tchouproff (13). For example noticing that his N is our P, and that his $\mu_4(N)=\mu_4(b)$ in our notation, we find from equation (8), p. 286 of his paper (13).

$$\mu_4(N)=\mu_4(b)=\mu_4(b)=\frac{3}{P^2}\cdot\mu^2[2,\ P]+\frac{1}{P^4}S_p[\mu_4(a_p)-3\mu_2{}^2(a_p)]$$

Since in Tchouproff's notation,

$$\mu[2,\ P]=\frac{1}{P}S_p[\mu_2(a_p)]$$

it immediately follows that

$$\begin{aligned}\mu_4(b)&=\frac{3}{P^4}S_p[\mu_2{}^2(a_p)]+\frac{6}{P^4}S_pS_r[\mu_2(a_p)\cdot\mu_2(a_r)]\\&\quad+\frac{1}{P^4}S_p[\mu_4(a_p)-3\mu_2{}^2(a_p)]\\&=\frac{6}{P^4}S_pS_r[\mu_2(a_p)\cdot\mu_2(a_r)]+\frac{1}{P^4}S_p[\mu_4(a_p)]\end{aligned}$$

which is identical with our equation (11.4).

12. To prevent confusion, we shall now restore the full notation, and write x_{pq}, $x_{pq}{}'$, $\bar{x}_{pq}$, $\bar{x}_{pq}{}'$ as the observed and "true" (or mean) values of the statistics for the pth character and the q and the q'th sample respectively. We shall also write $\Sigma_{pq}{}^2$, $\Sigma_{pq}{}'^2$ and $\Sigma y_p{}^2$ as the variances of x_{pq}, $x_{pq}{}'$ and $\bar{y}_p$.

Then $v_p{}^2=\dfrac{\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2}{(\bar{x}_{pq}-\bar{x}_{pq}{}')^2}$, $w_p{}^2=\dfrac{\Sigma y_p{}^2}{\bar{y}_p{}^2}$ .. .. (12·01).

If $\bar{y}_p=k_p=$constant, we may put $w_p{}^2=0$, and obtain

$$\bar{b}=\frac{1}{P}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}{}')^2}{k_p{}^2}\right]+\frac{1}{P}S_p\left[\frac{(\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2}{k_p{}^2}\right] \quad .. \quad (12{\cdot}1)$$

$$\begin{aligned}\mu_2(b)&=\frac{4}{P^2}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}{}')^2}{k_p{}^2}\cdot\frac{(\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2)}{k_p{}^2}\right]\\&\quad+\frac{2}{P^2}S_p\left[\frac{(\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2)^2}{k_p{}^4}\right] \quad .. \quad .. \quad (12{\cdot}2)\end{aligned}$$

$$\begin{aligned}\mu_3(b)&=\frac{24}{P^3}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}{}')^2}{k_p{}^2}\cdot\frac{(\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2)^2}{k_p{}^4}\right]\\&\quad+\frac{8}{P^3}S_p\left[\frac{(\Sigma_{pq}{}^2+\Sigma_{pq}{}'^2)^3}{k_p{}^6}\right] \quad .. \quad .. \quad (12{\cdot}3)\end{aligned}$$

$$\mu_4(b)=\frac{48}{P^4}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}')^4}{k_p^4}\cdot\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)^2}{k_p^4}\right]$$

$$+\frac{240}{P^4}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}')^2}{k_p^4}\cdot\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)^3}{k_p^6}\right]+\frac{60}{P^4}S_p\left[\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)^4}{k_p^8}\right]$$

$$-\frac{6}{P^4}S_pS_r\left[\left\{4\frac{(\bar{x}_{pq}+\bar{x}_{pq}')^2}{k_p^2}\cdot\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)}{k_p^2}\right.\right.$$

$$\left.+2\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)^2}{k_p^4}\right\}\left\{4\frac{(\bar{x}_{rq}+\bar{x}_{rq}')^2}{k_r^2}\cdot\frac{(\Sigma_{rq}^2+\Sigma_{rq}'^2)^2}{k_r^2}\right.$$

$$\left.\left.+2\frac{(\Sigma_{rq}^2+\Sigma_{rq}'^2)^2}{k_r^4}\right\}\right] \quad .. \quad .. \quad .. \quad (12{\cdot}4).$$

13. When the two samples are drawn from the same group or population $(\bar{x}_{pq}-\bar{x}_{pq}')=0$ for all values of p.

The mean value $(\bar{b})$ will not however vanish. We therefore introduce a small correcting term[1] and define any measure of group divergence by the general formula :—

$$U^2=\frac{1}{P}S_p\left[\frac{(x_{pq}-x_{pq}')^2}{k_p^2}\right]-\frac{1}{P}S_p\left[\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)}{k_p^2}\right] \quad .. \quad (13{\cdot}0)$$

so that the mean value of U^2 is given by

$$\bar{U}^2=\frac{1}{P}S_p\left[\frac{(\bar{x}_{pq}-\bar{x}_{pq}')^2}{k_p^2}\right] \quad .. \quad .. \quad .. \quad (13{\cdot}1).$$

It will be noticed that $\bar{U}^2=0$, when $(\bar{x}_{pq}-\bar{x}_{pq}')=0$ for all values of p, *i.e.*, for two samples drawn from the same group or population.

The variance and the other moment coefficient for U^2 will of course be the same as those for "b", and will be given by equations (11·1)–(11·4) or by equations (12·1)–(12·4) as the case may be.

---

[1] It is true that equation (13·0) may sometimes give a negative value of U^2 (or what amounts to the same thing, an imaginary value for U, the generalised distance between the two groups). It will be noticed, however, that the correcting term $\frac{1}{P}S_p\left[\frac{(\Sigma_{pq}^2+\Sigma_{pq}'^2)}{k_p^2}\right]$ is a quantity of the order of errors of random sampling, so that a negative value of U^2 will occur only when the observed value of the divergence is of the order of (or smaller than) the errors of random sampling. The statistical implication is obvious; in such cases the divergence must be treated as imaginary, *i.e.*, non-significant. It will be remembered that C^2, the Pearsonian Coefficient of Racial Likeness, will also (in similar circumstances) sometimes assume negative values.

IV. A Coefficient of Divergence in Means

14. We may now proceed to construct different coefficients of divergence by substituting suitable values for x_{pq}, x_{pq}', and k_p in equation (13.0).

Let us put

$$x_{pq}=m_{pq},\ x_{pq}'=m_{pq}',\ \text{and}\ k_p^2=\bar{\sigma}_p^2 \qquad .. \quad (14{\cdot}01)$$

where $\bar{\sigma}_p^2$ is a reliable constant value of the variance for the pth character.[1] If the size of the two samples are n_{pq}, n_{pq}' respectively, then (neglecting differences in variability between the two groups) we may write

$$\Sigma_{pq}^2=\frac{\bar{\sigma}_p^2}{n_{pq}},\ \Sigma_{pq}'^2=\frac{\bar{\sigma}_p^2}{n_{pq}'} \qquad .. \quad .. \quad (14{\cdot}02)$$

and $$v_p^2=\frac{\bar{\sigma}_p^2\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)}{(\bar{m}_{pq}-\bar{m}_{pq}')^2},\ w_p^2=0 \qquad .. \quad .. \quad (14{\cdot}03).$$

Calling this particular coefficient D^2, we have

$$D^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{1}{P}S_p\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \qquad .. \quad (14{\cdot}0).$$

With mean value

$$\bar{D}^2=\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\right] \qquad .. \quad .. \quad (14{\cdot}1)$$

$$\mu_2(D^2)=\frac{4}{P^2}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\cdot\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right]+\frac{2}{P^2}S_p\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^2\right] \qquad .. \quad (14{\cdot}2)$$

$$\mu_3(D^2)=\frac{24}{P^3}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')}{\bar{\sigma}_p^2}\cdot\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^2\right]+\frac{8}{P^3}S_p\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^3\right] \qquad .. \quad (14{\cdot}3)$$

---

[1] This constant value of $\bar{\sigma}_p^2$ may be taken from a very long series of measurements, or the average value of a fairly large number of estimates based on smaller samples may be used with advantage.

$$\mu^4(D^2)=\frac{48}{P^4}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^4}{\bar{\sigma}_p^4}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^2\right]$$
$$+\frac{240}{P^4}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\cdot\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^3\right]$$
$$+\frac{60}{P^4}S_p\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^4\right]$$
$$+\frac{6}{P^4}S_pS_r\left[\left\{4\,\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\cdot\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right.\right.$$
$$\left.+2\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^2\right\}\left\{4\,\frac{(\bar{m}_{rq}-\bar{m}_{rq}')^2}{\bar{\sigma}_r^2}\cdot\left(\frac{1}{n_{rq}}+\frac{1}{n_{rq}'}\right)\right.$$
$$\left.\left.+2\left(\frac{1}{n_{rq}}+\frac{1}{n_{rq}'}\right)^2\right\}\right] \quad .. \quad (14{\cdot}4)$$

where S_pS_r denotes a summation for all possible pairs of values of p and r, $(p\neq r)$.

15. If the size of the sample remains constant for all characters, *i.e.*, $n_{pq}=n_{rq}=\ldots n_q$, and $n_{pq}'=n'_{rq}=\ldots n_q'$, and we write,

$$\frac{2}{\bar{n}_q}=\left(\frac{1}{n_q}+\frac{1}{n_q'}\right) \quad .. \quad .. \quad .. \quad (15{\cdot}01)$$

then the above formulæ take a much simpler form.

Let us write

$$(\bar{d}_p^2)=\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2\cdot},\quad v_p^2=\frac{1}{\bar{d}_p^2}\cdot\frac{2}{\bar{n}_q} \quad .. \quad (15{\cdot}02)$$

and substitute these values in equations (10·61)-(10·91).

Then

$$\bar{a}_p=(\bar{d}_p^2)+\frac{2}{\bar{n}_q} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}1)$$

$$\mu_2(a_p)=\frac{8}{\bar{n}_q}\left(\bar{d}_p^2+\frac{1}{\bar{n}_q}\right) \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}2)$$

$$\mu_3(a_p)=\frac{32}{\bar{n}_q^2}\left(3\,\bar{d}_p^2+\frac{2}{\bar{n}_q}\right) \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}3)$$

$$\mu_4(a_p)=\frac{192}{\bar{n}_q^2}\left((\bar{d}_p^2)^2+10\,\frac{\bar{d_p^2}}{\bar{n}_q}+\frac{5}{\bar{n}_q^2}\right) \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}4).$$

Substituting these values in equations (11·1) to (11·4) we get

$$\bar{b}=\frac{1}{P}S_p[(\bar{d}_p^2)]+\frac{2}{\bar{n}_q} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}5)$$

$$\mu_2(b)=\frac{1}{P^2}S_p\left[\frac{8}{\bar{n}_q}\left(\bar{d}_p{}^2+\frac{1}{\bar{n}_q}\right)\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}6)$$

$$\mu_3(b)=\frac{1}{P^3}S_p\left[\frac{32}{\bar{n}_q{}^2}\left(3\,(\bar{d}_p{}^2)+\frac{2}{\bar{n}_q}\right)\right] \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}7).$$

Also choosing Tchouproff's form for $\mu_4(b)$, we have

$$\mu_4(b)=\frac{3}{P^2}\left\{\frac{1}{P}S_p[\mu_2(a_p)]\right\}^2+\frac{1}{P^4}S_p[\mu_4(a_p)-3\mu_2{}^2(a_p)]$$

Writing $\bar{D}^2=\frac{1}{P}S_p[(\bar{d}_p{}^2)] \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15{\cdot}03)$

we notice that

$$\left\{\frac{1}{P}S_p[\mu_2(a_p)]\right\}^2=\left\{\frac{1}{P}S_p\left[\frac{8}{\bar{n}q}\left(\bar{d}_p{}^2+\frac{1}{\bar{n}_q}\right)\right]\right\}^2=\left\{\frac{8}{\bar{n}_q}\left(\bar{D}^2+\frac{1}{\bar{n}_q}\right)\right\}^2$$

$$=\frac{64}{\bar{n}_q{}^2}\cdot\left\{(\bar{D}^2)+\frac{1}{\bar{n}_q}\right\}^2.$$

Again

$$\mu_4(a_p)-3\mu_2{}^2(a_p)=\frac{192}{\bar{n}_q{}^2}\left\{(\bar{d}_p{}^2)^2+10\,\frac{\bar{d}_p{}^2}{\bar{n}_q}+\frac{5}{\bar{n}_q{}^2}\right\}$$

$$-3\left\{\frac{8}{\bar{n}_q}\left(\bar{d}_p{}^2+\frac{1}{\bar{n}_q}\right)\right\}^2$$

$$=\frac{192\times4}{\bar{n}_q{}^3}\left\{2\,(\bar{d}_p{}^2)+\frac{1}{\bar{n}_q}\right\}$$

Therefore

$$\frac{1}{P^4}S_p[\mu_4(a_p)-3\mu_2{}^2(a_p)]=\frac{1}{P^4}S_p\left[\frac{192\times4}{\bar{n}_q{}^3}\left\{2(\bar{d}_p{}^2)+\frac{1}{\bar{n}_q}\right\}\right]$$

$$\| \frac{192\times4}{P^3\,.\,\bar{n}_q{}^3}\left[2(\bar{D}^2)+\frac{1}{\bar{n}_q}\right].$$

Thus

$$\mu_4(b)=\frac{3\times64}{P^2\,.\,\bar{n}_q{}^2}\left\{(\bar{D}^2)+\frac{1}{\bar{n}_q}\right\}^2+\frac{192\times4}{P^3\,.\,\bar{n}_q{}^3}\left\{2(\bar{D}^2)+\frac{1}{\bar{n}_q}\right\}$$

$$=\frac{192}{P^2\,.\,\bar{n}_q{}^2}\cdot\left[\left\{(\bar{D}^2)+\frac{1}{\bar{n}_q}\right\}^2+\frac{4}{P\,.\,\bar{n}_q}\left\{2(\bar{D}^2)+\frac{1}{\bar{n}_q}\right\}\right] \quad .. \quad (15{\cdot}8).$$

16. We may now sum up our results for the coefficient D^2 defined by

$$D^2\equiv\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq'})^2}{\bar{\sigma}_p{}^2}\right]-\frac{2}{\bar{n}_q} \quad .. \quad .. \quad (16{\cdot}0).$$

The mean value is given by

$$(\bar{D}^2)=\frac{1}{P}S_p\left[\frac{\overline{m}_{pq}-\overline{m}_{pq'})^2}{\bar{\sigma}_p{}^2}\right] \quad .. \quad .. \quad .. \quad (16{\cdot}1).$$

Writing $$\delta \equiv (\bar{n}_q \,.\, \bar{D}^2) \quad .. \quad .. \quad .. \quad .. \quad .. \quad (16{\cdot}11)$$

we have $$\mu_2(D^2) = \frac{8}{P \,.\, \bar{n}_q}\left[(\bar{D}^2) + \frac{1}{\bar{n}_q}\right] = \frac{8(\delta+1)}{P \,.\, \bar{n}_q{}^2} \quad .. \quad .. \quad (16{\cdot}2)$$

$$\mu_3(D^2) = \frac{32}{P^2 \,.\, \bar{n}_q{}^2}\left[3(\bar{D}^2) + \frac{1}{\bar{n}_q}\right] = \frac{32(3\delta+2)}{P^2 \,.\, \bar{n}_q{}^3} \quad .. \quad (16{\cdot}3)$$

$$\mu_4(D^2) = \frac{192}{P^2 \,.\, \bar{n}_q{}^2}\left[\left\{(\bar{D}^2) + \frac{2}{\bar{n}_q}\right\}^2 + \frac{4}{P \,.\, \bar{n}_q}\left\{2(\bar{D}^2) + \frac{1}{\bar{n}_q}\right\}\right]$$

$$= \frac{192}{P^2 \,.\, \bar{n}_q{}^4}\left[(\delta+1)^2 + \frac{4}{P}(2\delta+1)\right] \quad .. \quad .. \quad .. \quad (16{\cdot}4)$$

$$\beta_1 = \frac{2}{P}\frac{\left\{3(\bar{D}^2) + \frac{2}{\bar{n}_q}\right\}^2}{\left\{(\bar{D}^2) + \frac{1}{\bar{n}_q}\right\}^3} = \frac{2}{P} \cdot \frac{(3\delta+2)^2}{(\delta+1)^3} \quad .. \quad .. \quad .. \quad (16{\cdot}5)$$

$$\beta_2 = \frac{3\left[\left\{(\bar{D}^2) + \frac{1}{\bar{n}_q}\right\}^2 + \frac{4}{P \,.\, \bar{n}_q}\left\{2(\bar{D}^2) + \frac{1}{\bar{n}_q}\right\}\right]}{\left\{(\bar{D}^2) + \frac{1}{\bar{n}_q}\right\}^2}$$

$$= 3 + \frac{12}{P}\frac{(2\delta+1)}{(\delta+1)^2} = 3 + \frac{12\delta^2+18\delta+6}{9\delta^2+12\delta+4} \cdot \beta_1 \quad .. \quad .. \quad (16{\cdot}6).$$

Even when the size of the samples is not absolutely constant, the above formulæ may still be used without appreciable error if the fluctuation in the size of the sample is small, and we write

$$\frac{2}{\bar{n}_q} = \frac{1}{P} S_p\left[\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)\right] \quad .. \quad .. \quad (16{\cdot}7).$$

Further when the magnitude of (D^2) is of the order of, or greater than $\frac{10}{\bar{n}_q}$, it will usually be possible to neglect even large fluctuations in the size of the sample and use a mean value of $(\bar{n}_q)$ as defined in equation (16·7).

Finally when the two samples are drawn from the same population ($\bar{D}^2 = 0$), we have

$$(\bar{D}^2)_0 = O \pm \cdot 67449 \,.\, \frac{2}{\bar{n}_q}\sqrt{\frac{2}{P}} \quad .. \quad .. \quad (16{\cdot}8).$$

17. Since the standard deviation of (D^2)

$$= \frac{1}{\bar{n}_q}\sqrt{\frac{8(\delta+1)}{P}}$$

we notice that the ratio

$$\frac{\text{Value of } D^2}{\text{Standard deviation of } D^2} = \frac{\delta}{\bar{n}_q} \cdot \bar{n}_q \sqrt{\frac{P}{8(\delta+1)}}$$

$$= \sqrt{\frac{P \cdot \delta^2}{8(\delta+1)}} \equiv e \ldots \quad (17{\cdot}1).$$

For any assigned value of e, that is for any given standard of statistical significance, the above equation furnishes a numerical relation between P and $\delta(=\bar{n}_q \cdot \bar{D}^2.)$

For example, if we decide to consider D^2 to be significantly different from zero when the numerical value of D^2 exceeds 2·5 times the standard deviations of D^2, that is if we fix the level of significance at $e=2.5$ (which corresponds roughly to odds of 80 to 1 in the case of a normal distribution), we get

$$P = \frac{50(\delta+1)}{\delta^2} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (17.2).$$

For moderately large values of δ, P is approximately equal to 50 δ, or $P.\delta = P.\bar{n}_q \cdot \bar{D^2} = 50$ approximately.

For any given value of P, equation (17·2) may also be used to determine the lower limit of δ for which divergence can be asserted with safety. For example for $P=1$, δ must be greater than 50; for $P=10$, δ must be greater than 6: and for $P=20$, δ must not be less than 3. In usual anthropological practice it will not be often possible to increase P beyond 20. and almost never beyond 100. We conclude therefore that even under the most favourable circumstances ($P=100$ or more) the size of the sample ($\bar{n}_q$) must be large enough to yield a value of δ greater than 1, while usually (for $P=20$ approximately) the value of δ must be greater than 3 or 4.

18. We may now investigate the nature of the frequency distribution of (D^2).

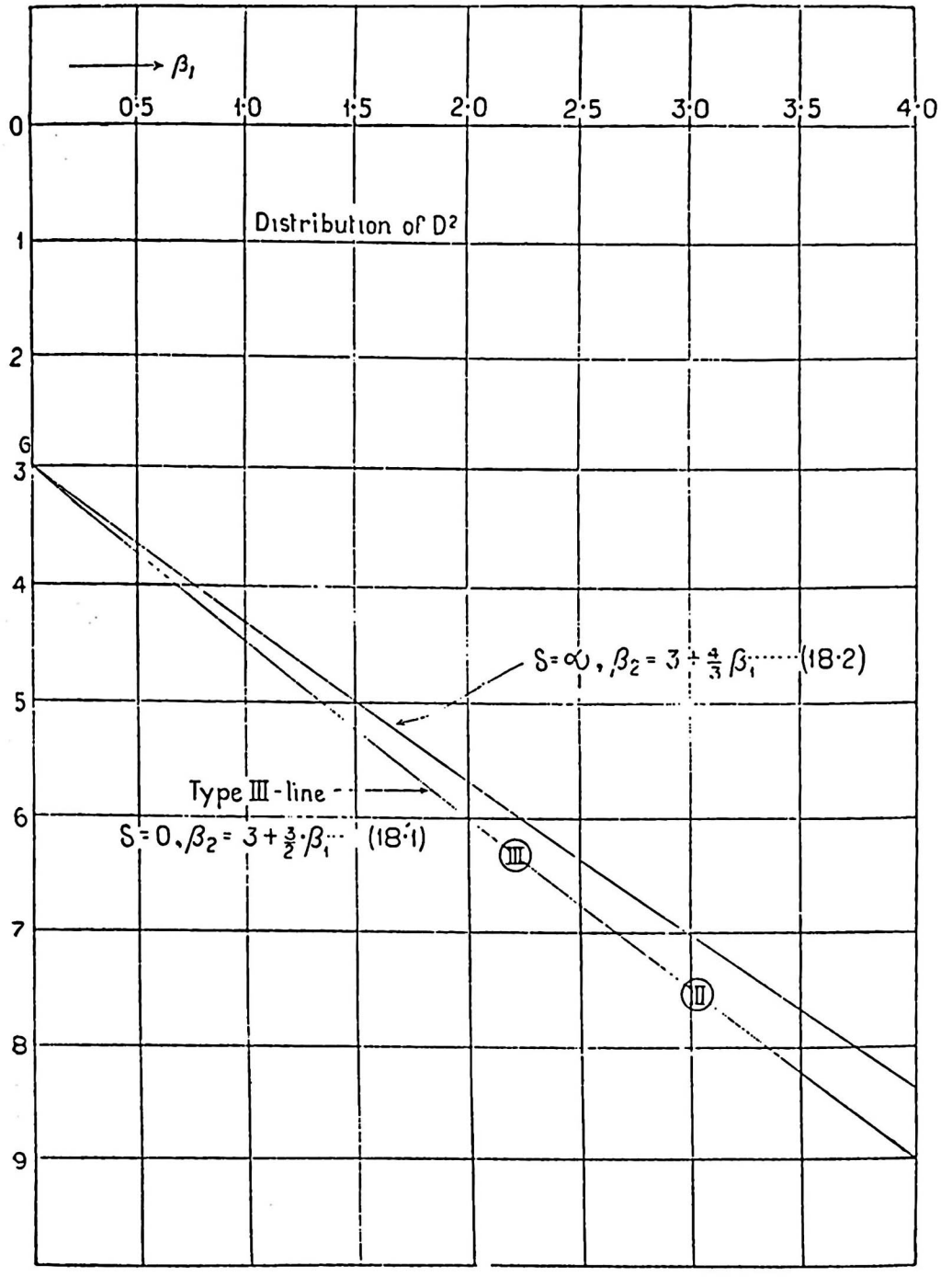
$$\text{For } \delta=0,\ \beta_2 = 3 + \tfrac{3}{2}\beta_1 \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (18.1).$$

The Pearsonian criterion $k_1 = (6 + 3\beta_1 - 2\beta_2) = 0$, and the distribution will belong to Type III of the Pearsonian family of curves.

$$\text{Again for } \delta = \infty,\ \beta_2 = 3 + \tfrac{1}{3} \cdot \beta_1 \ldots \quad \ldots \quad \ldots \quad \ldots \quad (18.2).$$

It will be easily seen from the accompanying sketch that equation (18·2) gives a straight line lying wholly in the Type I region on the $\beta_2 - \beta_1$ diagram.

We conclude therefore that the distribution of D^2 will conform generally to Type I of the Pearsonian family, except in the case of two groups (or samples) taken from the same population, when the distribution will pass into the Type III curve.



"On Tests and Measures of Group Divergence."

19. When $\delta=\overline{D}^2=0$, *i.e.*, the two samples belong to the same population, we have (1, p . 90)

$$\left.\begin{aligned} &D^2=0,\ \mu_2=\frac{8}{P.\ \bar{n}_q{}^2},\ \mu_3=\frac{64}{P^2\ .\ \bar{n}_q{}^3},\ \mu_4=\frac{192}{P^2\ .\ \bar{n}_q{}^4}\left(1+\frac{4}{P}\right)\\ &\beta_1=\frac{8}{P},\ \beta_2=3+\frac{12}{P}\end{aligned}\right\} \quad .. \quad (19.1).$$

The equation to the frequency curve is given by

$$y=y_0\ .\ e^{-\frac{px}{a}}\ .\left(1+\frac{x}{a}\right)^p \text{ with origin at mode} \quad .. \quad (19.2)$$

$$\left.\begin{aligned} &\text{where } p=\frac{4}{\beta_1}-1=\frac{P}{2}-1,\\ &a=\frac{2\mu_2{}^2}{\mu_3}-\frac{\mu_3}{2\mu_2}=\frac{2}{\bar{n}_q}\left(1-\frac{2}{P}\right)\\ &y_0=\frac{N}{a}\ .\ \frac{p^{(p+1)}}{e^p\ .\ \Gamma(p+1)}\\ &\text{Mode}-\text{Mean}=-\tfrac{1}{2}\frac{\mu_3}{\mu^2}=-\frac{4}{P\bar{n}_q}\\ &\text{Start of the curve}=\{\text{Mode}-\text{“}a\text{”}\}=-\frac{2}{\bar{n}_q}.\end{aligned}\right\} \quad .. \quad (19{\cdot}3).$$

When δ is small in comparison with 1, *i.e.*, the two samples belong to closely associated groups, we may still use a Type III curve without serious error. In ascending power of δ we have

$$\left.\begin{aligned} &\beta_1=\frac{8}{P}\ (1-\tfrac{3}{4}\delta^2+\tfrac{5}{4}\delta^3-\tfrac{3}{2}\delta^4)\\ &\beta_2=3+\frac{12}{P}\ (1-\delta^2+2\delta^3-3\delta^4)\\ &p=\frac{P}{2}\ (1+\tfrac{3}{4}\delta^2-\tfrac{5}{4}\delta^3+\tfrac{33}{16}\ .\ \delta^4)-1\\ &a=\frac{2}{\bar{n}_q}\ (1+\tfrac{1}{2}\delta+\tfrac{1}{4}\delta^2-\tfrac{3}{8}\delta^3+\tfrac{9}{16}\delta^4)\\ &\qquad\qquad -\frac{4}{P\ .\ \bar{n}_q}\ (1+\tfrac{1}{2}\delta-\tfrac{1}{2}\delta^2+\tfrac{1}{2}\delta^3-\tfrac{1}{2}\delta^4)\\ &\text{Mode}-\text{Mean}=-\frac{4}{P\ .\ \bar{n}_q}(1+\tfrac{1}{2}\delta-\tfrac{1}{2}\delta^2+\tfrac{1}{2}\delta^3-\tfrac{1}{2}\delta^4)\\ &\text{Start of the curve}=-\frac{2}{\bar{n}_q}\ (1+\tfrac{1}{4}\delta^2-\tfrac{3}{8}\delta^3+\tfrac{9}{16}\delta^4)\end{aligned}\right\} \quad .. \quad (19{\cdot}4).$$

When δ is significantly different from zero, we have a Type I curve defined by (1, *p* . 54):—

$$y=y_0\left(1+\frac{x}{a_1}\right)^{m_1}.\left(1+\frac{x}{a_2}\right)^{m_2} \quad .. \quad .. \quad .. \quad .. \quad (19{\cdot}5)$$

where the frequency constants are given by the following equations.

$$\text{Let } r\equiv\frac{6(\beta_2-\beta_1-1)}{6+3\beta_1-2\beta_2}=\left\{6+\frac{4(3\delta+1)}{\delta^2}\right\}+\frac{2(\delta+1)^3}{\delta^2}\,.\,P \ldots \quad (19{\cdot}61)$$

$$\text{and } g^2=16(r+1)+\beta_1(r+2)^2$$

$$=\left[\frac{8(\delta+1)^3(13\delta^2+12\delta+4)}{\delta^4}\,.\,P+\right.$$

$$+\left\{688+\frac{32(57\delta^3+55\,.\,\delta^2+24\delta+4)}{\delta^4}\right\}$$

$$\left.+\frac{32(3\delta+2)^2(4\delta^4+12\delta^3+13\delta^2+6\delta+1)}{(\delta+1)^3\,.\,\delta^4\,.\,P}\right] \ldots \quad (19{\cdot}62).$$

$$\left.\begin{aligned}&\text{Then } a_1+a_2=\tfrac{1}{2}g\sqrt{\mu_2}, \quad \frac{m_1}{a_1}=\frac{m_2}{a_2}\\&(m_1\,,\,m_2)=\tfrac{1}{2}(r-2)\pm\tfrac{1}{2}\frac{r(r-2)}{g}\sqrt{\beta_1}\\&y_0=\frac{N}{(a_1+a_2)}\cdot\frac{(m_1)^{m_1}\,.\,(m_2)^{m_2}}{(m_1+m_2)^{m_1+m_2}}\cdot\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}\\&\text{Mode}-\text{Mean}=-\tfrac{1}{2}\frac{\mu^3}{\mu^2}\,.\left(\frac{r+2}{r-2}\right)\end{aligned}\right\} \ldots \quad (19{\cdot}7).$$

When δ is large in comparison with 1, expanding in powers of $\left(\frac{1}{\delta}\right)$,

$$\left.\begin{aligned}&\beta_1=\frac{18}{P\,.\,\delta}\left\{1-\frac{5}{3}\left(\frac{1}{\delta}\right)+\frac{22}{9}\left(\frac{1}{\delta}\right)^2-\frac{10}{3}\left(\frac{1}{\delta}\right)^3+\frac{13}{3}\left(\frac{1}{\delta}\right)^4\right\}\\&\beta_2=3+\frac{24}{P\,.\,\delta}\left[1-\frac{3}{2}\left(\frac{1}{\delta}\right)\left\{1-\left(\frac{41}{3\delta}\right)\right.\right.\\&\qquad\left.\left.+\frac{5}{3}\left(\frac{1}{\delta}\right)^2-\frac{6}{3}\left(\frac{1}{\delta}\right)^3\right\}\right]\\&r=\left\{6+12\left(\frac{1}{\delta}\right)+4\left(\frac{1}{\delta}\right)^2\right\}\\&\qquad+2\left\{(\delta+3)+3\left(\frac{1}{\delta}\right)+\left(\frac{1}{\delta}\right)^2\right\}\,.\,P\end{aligned}\right\} \ldots \quad (19{\cdot}8).$$

V. Results of Sampling Experiments

20. I give below the results of a few sampling experiments which were undertaken to test the above formulæ.

Remembering that the original material (from which the samples were drawn) was supposed to obey the normal law of frequency, we can use a table of the probability integral (Biometric Table II, pp. 2–9) together with Tippett's "Random Sampling Numbers" (Tracts for Computers No. XV) in the manner described in illustration I, page (iv) of the above tract, to form samples from a normal population. Using 11 sheets (Nos. 1, and 16–25 of the above tract), $11\times400=4{,}400$ individual random samples from a normal population were obtained. Combining 5 such samples at a time, $11\times80=880$ independent samples of the mean of 5 individuals were next calculated. Combining these 880 values in different ways and subtracting, 4000 sampled values of $(m-m')$, (or rather of $(m-m'/\sigma)$, since the sampled values were all expressed in terms of their standard deviation) were obtained. Squaring such differences, I finally obtained 4000 sampled values of $\frac{(m-m')^2}{\sigma^2}$.

Taking them separately we have obviously a sample of 4000 values of D^2 with $\bar{n}_q=5$, $P=1$, and $\bar{D}^2=0$, (on the assumption that all the samples are truly random[1]). Again adding them up in batches of 5, 10, and 20, I obtained samples of size $N=800$, 400, 200, and $P=5$, 10, 20 respectively. The sampled values were then grouped, and the frequency constants calculated in the usual way. The actual mean value was calculated by direct addition (without grouping) in order to keep it free from errors due to grouping. The mean value obtained from the grouped figures have been given within a square bracket only for purposes of comparison.

For $\bar{D}^2=0$, $\bar{n}_q=5$, we have

$$\left.\begin{aligned}&\mu_2=\frac{0\cdot32}{P},\ \mu_3=\frac{0\cdot512}{P^2},\ \mu_4=\frac{0\cdot3072}{P^2}\left(1+\frac{4}{P}\right)\\&\beta_1=\frac{8}{P},\ \beta_2=3\left(1+\frac{4}{P}\right),\ \text{Mode}-\text{mean}=-\frac{0\cdot8}{P}\end{aligned}\right\}\quad\ldots\quad\ldots\quad(20\cdot1)$$

(1) $P=1$, $\bar{n}_q=5$, $\bar{D}^2=0$, $\beta_1=8$, $\beta_2=15$, $N=4000$, $\chi_1=\cdot010665$.

---

[1] On more careful consideration I am inclined to think that this assumption was not strictly fulfilled in my experiments, for the reason that only 880 independent values of m(or m') were used to obtain 4000 values of $(m-m')$ so that a certain amount of repetition was inevitable. This point has been further discussed for $\bar{n}_q=20$.

Using equations (lii), (liii), and (liv) on page lxi of the Introduction to the Tables for Statisticians (10), I find

$$\beta_3=272,\ \beta_4=755,\ \beta_5=19,\ 752,\ \beta_6=74,\ 417.$$

Using equations (lxxv) and (lxxv bis) on p. lxv of the same Introduction we get

$$\sqrt{N}\Sigma_{\beta_1}=78{\cdot}97,\ \sqrt{N}\Sigma_{\beta_2}=205{\cdot}6.$$

Again using the same values of the β-constants in equation (27) of Kazutaro Yasukawa's paper "On the Probable Error of the Mode of Skew Frequency Distribution" (17, p. 266) I found $B_1=-125/18$, $B_2=-7/6$. $B_3=+19/6$, and $B_4=-1\,9$. Substituting these values in equation (29) of the same paper I obtained the ratio of the probable error of the mode to the probable error of the mean=163 approximately.

TABLE 1

Statistics	Expected	Observed	Difference
Mean	0	+0·0208	0·0208 ± 0·0060
Mode--Mean	−0·8	−0·7470	0·0530 ± 0·9834
μ_2	0·32	0·3158	0·0042 ± 0·0128
β_1	8·00	7·0675	0 9325 ± 0·8421
β_2	15·00	12·6000	2·4000 ± 2·0574

The probable error was of course calculated from the expected value of the constant in each case. A glance at column 4 will show that agreement between expected and observed values is quite satisfactory.

(2) $P=5$, $\bar{n}_q=5$, $\bar{D}^2=0$, $\beta_1=1{\cdot}6$, $\beta_2=5{\cdot}4$, $N=800$, $\chi_1={\cdot}02385$. Following the same procedure I found

$$\beta_3=23{\cdot}68,\ \beta_4=86{\cdot}2,\ \beta_5=555{\cdot}84,\ \beta_6=2548{\cdot}84$$

and $B_1=-53/28$, $B_2=+5/14$, $B_3=+23\ 14$, $B_4=-5/21$

Hence $\sqrt{N}\Sigma_{\beta_1}=12{\cdot}70$, $\sqrt{N}\Sigma_{\beta_2}=35{\cdot}52$, and the ratio of the probable error of the mode to the probable error of the mean= 4·775.

We now have the following table:—

TABLE 2

Statistics	Expected	Observed	Difference ± Probable error
Mean	0	[0·0190]	0·0208 ± 0·00803
Mode—Mean	−0·16	−0·1593	0·0007 ± 0·0288
μ_2	0·064	0·0663	0·0023 ± 0·0032
β_1	1·6	1·5303	0·0697 ± 0·3038
β_2	5·4	5·2451	0·1549 ± 0·7942

(3) $P=10$, $\bar{n}_q=5$, $D^2=0$, $\beta_1=0{\cdot}8$, $\beta_2=4{\cdot}2$, $N=400$, $\chi_1=\cdot03372$.

From Yasukawa (p. 277, Table II) I find ratio of probable error of mode to the probable error of mean=2·3991.

TABLE 3

Statistics	Expected	Observed	Difference
Mean	0	[0·0208]	0·0208 ± ·0060
Mode−mean	−0·08	−0·0777	·0023 ± ·0145
μ_2	0·0320	0·0355	·0035 ± ·0019
β_1	0·8	1·0919	·2919 ± ·2293
β_2	4·2	4·5253	·3253 ± ·6811

(4) $D^2=0$, $\bar{n}_q=5$, $P=20$, $N=200$, $\chi_1=\cdot04769$.

TABLE 4

Statistics	Expected	Observed	Difference
Mean	0	[0·0213]	·0208 ± ·00603
Mode-mean	−0·04	−0·0664	·0264 ± ·0105
μ_2	0·016	0·017535	·001536 ± ·001230
β_1	0·4	1·0360	·6360 ± ·1908
β_2	3·6	4·6651	1·0651 ± ·6176

21. Following the same procedure and using the same converted sheets (Nos. 1, 16-25) of tract No. XV (15), 220 samples of means of 20 (*i.e.*, $\bar{n}_q=20$) were obtained, combining them in different ways, and squaring 4,000 sampled values of $(m-m')^2/\sigma^2$ for $n_q=20$, $P=1$, and $\overline{D}^2=0$ (on the assumption of random sampling) were calculated. In the present example we have :

$$D^2=0,\ \bar{n}_q=20,\ \mu_2=\frac{0{\cdot}02}{P},\ \mu_3=\frac{0{\cdot}008}{P^2},\ \mu_4=\frac{\cdot0012}{P^2}\left(1+\frac{4}{P}\right)$$

$$\beta_1=\frac{8}{P},\ \beta_2=3+\frac{12}{P},\ \text{Mode-mean}=-\frac{0{\cdot}2}{P}$$

(5) $\overline{D}^2=0$, $\bar{n}_q=20$, $P=1$, $N=4{,}000$, $\chi_1=\cdot010665$.

TABLE 5

Statistics	Expected	Observed	Difference
Mean	0	[·0163]	·012080 ± ·001508
Mode-mean	−0·2	−·1998	·0002 ± ·2445
μ_2	·02	·022720	·002720 ± ·000798
β_1	8·00	7·0277	·9723 ± ·8422
β_2	15·00	12·7096	2·2904 ± 2·0574

The mean value of $\bar{D}^2$ as directly calculated (without grouping) was 0·01208, and this is the value shown in the Table. The grouped value has been shown within square brackets.

(6) $\bar{D}^2=0$, $\bar{n}_q=20$, $P=5$, $N=800$, $\chi_1=\cdot02385$.

TABLE 6

Statistics	Expected	Observed	Difference
Mode-mean	−·04	·03474	·00526 ± ·00720
μ_2	·0040	·004073	·000073 ± ·000200
β_1	1·60	1·1853	·4147 ± ·3038
β_2	5·40	4·7273	·6727 ± ·7942

(7) $\bar{D}^2=0$, $\bar{n}_q=20$, $P=10$, $N=400$, $\chi_1=\cdot03372$.

TABLE 7

Statistics	Expected	Observed	Difference
Mode-mean	−0·01	−0·017183	·002807 ± ·001508
μ_2	0·002	·002066	·000060 ± ·003618
β_1	0·8000	·5734	·2266 ± ·2293
β_2	4·2000	4·5795	·3795 ± ·6811

(8) $\bar{D}^2=0$, $\bar{n}_q=20$, $P=20$, $N=200$, $\chi_1=\cdot04769$.

TABLE 8

Statistics	Expected	Observed	Difference
Mode-mean	−0·01	−0·012191	·002191 ± ·002637
μ_2	·0010	·001346	·000346 ± ·000077
β_1	·4000	·4416	·0416 ± ·1903
β_2	3·6000	2·8851	·7149 ± ·6176

The agreement with theory is satisfactory in every case with the single exception of the mean value of $\bar{D}^2$. Instead of the expected value $\bar{D}^2=0$, we actually obtain $\bar{D}^2=0·012\pm·0015$ showing a deviation of 8 times its probable error. As this discrepancy was very puzzling, I checked the whole arithmetic most carefully, but without any tangible results. On more careful consideration I am inclined to think, that the discrepancy may be attributed to a slight bias (or deviation from random sampling) introduced at the stage of obtaining the differences $(m-m')$. It will be remembered that from the 11

converted sheets we had $11 \times 20 = 220$ independent samples of means of 20. These 220 values were used over and over again (although always in different combinations) to yield 4,000 values of $(m - m')$. A bias was created owing to the fact that the different combinations were taken in a certain systematic order, and not in a perfectly random manner, and also because the process was stopped when the number of differences reached 4,000, so that all possible combinations could not be included.

I now realise that I ought to have (*a*) used a larger number of Tippet's sheets, and (*b*) formed 8,000 random values of means of 20 before proceeding to take differences. I intend to repeat the experiment at the earliest opportunity.

If we assume that owing to the bias discussed above the actual value of $\bar{D}^2 = \cdot 012$ (and not zero), with $\bar{n}_q = 20$, $\cdot\delta = 0 \cdot 24$, we have

$$\mu_2 = \frac{\cdot 0228}{P}, \quad \mu_3 = \frac{\cdot 01088}{P^2}, \quad \mu_4 = \frac{\cdot 0012}{P^2}\left[1 \cdot 5376 + \frac{5 \cdot 92}{P}\right]$$

$$\beta_1 = \frac{7 \cdot 7429}{P}, \quad \beta_2 = 3 + \frac{11 \cdot 5547}{P},$$

which lead to a slight improvement in the agreement with expected values.

22. I next turned to the Type I curve for values of $\bar{D}^2 \neq 0$, *i.e.*, for samples drawn from different groups or populations.

(9) Taking one series of 800 values of $(m - m')$ I found that the mean value of $\bar{D}^2$ was 0·00 9564. Now adding 0·2 to each individual value of $(m - m')$, and squaring, I obtained a sample of 800 for $\bar{D}^2 = \cdot 0095\ 64 + (0 \cdot 2)^2 = 0 \cdot 049564$. Calculating the frequency constants in the usual way I got the following results:—

$\bar{D}^2 = 0 \cdot 049564$, $\bar{n}_q = 20$, $\delta = 0 \cdot 99128$, $P = 1$, $N = 800$, $\chi_1 = \cdot 02385$.

TABLE 9

Statistics	Expected	Observed	Difference
Mean	·049564	·055416	·005852 ± ·004760
μ_2	·039826	·037837	·001989 ± ·003152
β_1	6·2663	5·1978	1·0685 ± 1·2290
β_2	12·0262	9·6168	2·4094 ± 2·7835

(10) Taking a second series of 800 (for which $\bar{D}^2$ was ·004050) adding 0·8, and squaring, I get a second sample for

$\bar{D}^2 = \cdot 644050$, $\delta = 12 \cdot 88$, $\bar{n}_q = 20$, $P = 1$, $N = 800$, $\chi_1 = \cdot 02385$.

TABLE 10

Statistics	Expected	Observed	Difference
Mean	·644050	·663126	·019076 ± ·012564
μ_2	·277620	·280092	·002472 ± ·012678
β_1	1·2352	·9696	·2656 ± ·2005
β_2	4·6044	3·9359	·6685 ± ·5020

(11) In the same way, adding 1·1 each to a sample of 800 values of $(m-m')$ for which $(\bar{D}^2)$ was ·005049, I obtained a sample for

$\bar{D}^2=1{\cdot}215049$, $\delta=25{\cdot}3$, $\bar{n}_q=20$, $P=1$, $N=800$.

TABLE 11

Statistics	Expected	Observed	Difference
Mean	1·215049	1·182362	·032687 ± ·016625
μ_2	·486020	·563690	·07767 ± ·020928
β_1	·6928	1·3798	·6870 ± ·1331
β_2	3·9299	5·0756	1·1456 ± ·3726

As a last example I added all the above 3 sets of 800 each, and taking the average of each triplet obtained a sample for

$\bar{D}^2=0{\cdot}636221$, $\delta=12{\cdot}7244$, $\bar{n}_q=20$, $P=3$, $N=800$.

TABLE 12

Statistics	Expected	Observed	Difference
Mean	·636221	·633634	·002587 ± ·007188
μ_2	·090830	·084744	·006086 ± ·003506
β_1	·6242	·4730	·1512 ± ·1009
β_2	3·5617	3·3765.	·1852 ± ·2455

23. We have thus tested experimentally the distribution for $\bar{n}_q=5$ and 20, $\delta=0$, and $P=1$, 5, 10, and 20. We have also tested the distribution for $\bar{n}_q=20$, $P=1$, and $\delta=0{\cdot}8$, $12{\cdot}8$, $25{\cdot}2$ (approximately), and finally for $\delta=12{\cdot}72\ 44$, $\bar{n}_q=20$, $P=3$.

The difference between expected and observed values of the frequency constants was in most cases less than twice the corresponding probable error. In one case ($\bar{D}^2=0$, $\bar{n}_q=20$) the mean value gave a highly discrepant result. We have reasons for believing, however, that this may be attributed to a bias introduced at a certain stage of the sampling experiment. In one other case ($\bar{n}_q=20$, $P=1$, $\bar{D}^2=1{\cdot}215$) the agreement is not

good, but taking the results as a whole they may be considered quite satisfactory.[1]

VI. Other Coefficients of Divergence in Means

24. We can construct other measures of divergence in means by choosing different values for k_p. Let us take $k_p^2=\bar{s}_p^2$, where $\bar{s}_p^2$ is a reliable *constant* value of the inter-class variance. Then we obtain a second coefficient of divergence in means

$$D_2^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\bar{s}_p^2}\right]-\frac{1}{P}S_p\left[\frac{\bar{\sigma}_p^2}{\bar{s}_p^2}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad .. \quad (26{\cdot}0)$$

with mean value

$$(\bar{D}_2^2)=\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{s}_p^2}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (26{\cdot}1)$$

and

$$\mu_2(D_2^2)=\frac{4}{P^2}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{s}_p^2}\left\{\frac{\bar{\sigma}_p^2}{\bar{s}_p^2}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right\}\right]$$
$$+\frac{2}{P^2}S_p\left[\frac{\bar{\sigma}_p^4}{\bar{s}_p^4}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)^2\right] \quad .. \quad (26{\cdot}2).$$

If the size of the samples remains constant[2] for all characters, we may write as before $\frac{2}{\bar{n}_q}=\left(\frac{1}{n_q}+\frac{1}{n_q'}\right)$, and obtain

$$D_2^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m'_{pq})^2}{\bar{s}_p^2}\right]-\frac{2}{\bar{n}_q}\cdot\frac{1}{P}S_p\left[\left(\frac{\bar{\sigma}_p^2}{\bar{s}_p^2}\right)\right] \quad .. \quad (26{\cdot}3)$$

$$(D_2^2)=\frac{1}{P}S\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{s}_p^2}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (26{\cdot}4)$$

$$\mu_2(D_2^2)=\frac{8}{P^2\cdot\bar{n}_q}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{s}_p^2}\cdot\frac{\bar{\sigma}_p^2}{\bar{s}_p^2}\right]$$
$$+\frac{8}{P^2\cdot\bar{n}_q^2}\cdot S_p\left[\left(\frac{\bar{\sigma}_p^4}{\bar{s}_p^4}\right)\right] \quad .. \quad (26{\cdot}5).$$

It is also possible to derive an exactly similar set of equations for a third coefficient D_3^2 by putting $k_p^2=\Sigma_p^2$, where Σ_p^2 is a

---

[1] I am indebted to my assistant Mr. Sudhir Kumar Banerjee for help in the arithmetical calculations in the sampling experiments.

[2] Or when the fluctuation in the size of the sample can be neglected, and a mean value of $\bar{n}_q$ as defined by equation (16·7) can be used without appreciable error.

reliable constant value of the "familial" variance as defined in equation (2·8).

25. It will be noticed that we have deduced the above expressions on the assumption that $\bar{\sigma}_p^2$, $\bar{s}_p^2$ (or Σ_p^2) are all constants, that is, on the assumption that the variance of these quantities are negligibly small. This assumption will be justified only when the estimates are based on a very large number of individual observations. In actual practice it would some time happen that we are obliged to base the estimates for $\bar{\sigma}_p^2$, $\bar{s}_p^2$, or Σ_p^2 on observed samples. In such cases, unless the size of the sample is very large, it would not be proper to neglect the variance of these quantities.

Let us consider the case $x_{pq}=m_{pq}$, $x_{pq}'=m_{pq}'$, and $k_p^2=\sigma_p^2$, where σ_p^2 is an estimate of variance based on n_p *effective* observations. If the size of the two samples are n_{pq}, n_{pq}' respectively, we have

$$v_p^2=\frac{\bar{\sigma}_p^2\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)}{(\bar{m}_{pq}-\bar{m}_{pq}')^2} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (27\cdot1)$$

$$w_p^2=\frac{\sigma_p^2}{2n_p}\cdot\frac{1}{\bar{\sigma}_p^2}=\frac{1}{2n_p} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (27\cdot2).$$

We then have for the quantity

$$b=\frac{1}{P}\cdot S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\sigma_p^2}\right] .. \quad .. \quad .. \quad .. \quad .. \quad (27\cdot3)$$

the mean value

$$\bar{b}=\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\left\{1+\frac{\bar{\sigma}_p^2\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)}{(\bar{m}_{pq}-\bar{m}_{pq}')^2}\right\}(1+\alpha_p)\right] \quad (27\cdot4)$$

$$=\frac{1}{P}S_p\left[\frac{\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}(1+\alpha_p)\right]$$

$$+\frac{1}{P}S\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)(1+\alpha_p)\right] \quad .. \quad (27\cdot41)$$

where

$$\alpha_p\equiv\left(\frac{3}{2\,.\,n_p}+\frac{15}{4\,.\,n_p^2}+\frac{105}{8\,.\,n_p^3}+\frac{945}{16\,.\,n_p^4}+\frac{10{,}395}{32\,.\,n_p^5}\right) \quad .. \quad (27\cdot5)$$

is a purely numerical factor.

When $n_p=n$ is same for all characters, and n_{pq}, n_{pq}' are also constant, so that we may write $\frac{2}{n_q}=\left(\frac{1}{n_q}+\frac{1}{n_q'}\right)$, we can define a measure of divergence

$$D_1^2=\frac{1}{(1+\alpha_1)}\cdot\frac{1}{P}S\left[\frac{(m_{pq}-m_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{2}{\bar{n}_q} \quad .. \quad .. \quad (27{\cdot}6)$$

where

$$\alpha_1=\left(\frac{3}{2n}+\frac{15}{4n^2}+\frac{105}{8n^3}+\frac{945}{16n^4}+\frac{10{,}395}{32\,.\,n^5}\right) \quad .. \quad .. \quad .. \quad (27{\cdot}51).$$

When n_{pq}, n_{pq}', and n_p are not constant, we may still define

$$D_1^2=\frac{1}{P}S_p\left[\left\{\left(\frac{1}{1+\alpha_p}\right)\frac{(m_{pq}-m_{pq}')^2}{\bar{\sigma}_p^2}\right\}-\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad (27{\cdot}61)$$

where α_p is given by equation (27·5).

The mean value is given in both cases by

$$(\bar{D}_1^2)=\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{\bar{\sigma}_p^2}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (27{\cdot}7)$$

which will again vanish when both the samples are drawn from the same population.

From equations (11·2) and (10·7) we also have

$$\begin{aligned}\mu_2(D_1^2)&=\frac{\alpha_2'}{(1+\alpha_1)^2}\left(\frac{8}{P.\bar{n}_q}\right)\left[(\bar{D}_1^2)+\frac{1}{\bar{n}_q}\right]\\&\qquad+\frac{2\alpha_3'}{(1+\alpha_1)^2}\cdot\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^4}{\bar{\sigma}_p^4}\right]\\&=(1+\alpha_2)\cdot\frac{8}{P\,.\,\bar{n}_q}\left[(\bar{D}_1^2)+\frac{1}{\bar{n}_q}\right]\\&\qquad+2\alpha_3\cdot\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^4}{\bar{\sigma}_p^4}\right] \quad .. \quad .. \quad (27{\cdot}8)\end{aligned}$$

where

$$\left.\begin{aligned}\alpha_2'&\equiv 1+\frac{1}{2n}+\frac{138}{4n^2}+\frac{1{,}740}{8\,.\,n^3}+\frac{24{,}615}{16\,.\,n^4}\\ \alpha_3'&=\frac{2}{2.n}+\frac{54}{4.n^2}+\frac{558}{8.n^3}+\frac{8{,}526}{16.n^4}\end{aligned}\right\}$$

and

$$\left.\begin{aligned}\alpha_2&=\frac{3}{n}+\frac{63}{4\,.\,n^2}+\frac{207}{3.n^3}+\frac{12{,}645}{16.n^4}\\ \alpha_3&=\frac{1}{n}+\frac{27}{2\,.\,n^2}+\frac{279}{4\,.\,n^3}+\frac{4{,}263}{8.n^4}\end{aligned}\right\} \quad .. \quad .. \quad (27{\cdot}9).$$

It will be noticed from the above formulæ that n must be fairly large in order that α_1, α_2, and α_3 may be negligibly small. If this condition is not fulfilled, and values of σ_p^2 based on small samples are used for calculating D^2, we should not be surprised if considerable fluctuations occur in observed values of D^2 from sample to sample.

26. Again let us choose $k_p^2 = M_p^2$, where M_p^2 is a reliable value of the inter-class mean for the pth character. As before we have

$$v_p^2 = \frac{\bar{\sigma}_p^2\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)}{(\bar{m}_{pq} - \bar{m}_{pq'})^2} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (28{\cdot}01)$$

and

$$w_p^2 = \frac{s_p^2}{M_p^2} \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (28{\cdot}02)$$

and we can easily obtain the necessary formulæ by substituting these values in equations (10·6)–(10·9) and (11·1)–(11·4).

If M_p is derived from wider material than the samples under consideration, we may treat it as a constant and put $w_p^2 = 0$. In this case we have

$$D_4^2 = \frac{1}{P} S_p\left[\frac{(m_{pq} - m_{p'q})^2}{M_p^2}\right] - \frac{1}{P} S_p\left[\frac{\bar{\sigma}_p^2}{M_p^2}\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)\right] .. \quad (28{\cdot}1)$$

with mean value

$$\bar{D}_4^2 = \frac{1}{P} S_p\left[\frac{(\bar{m}_{pq} - \bar{m}_{pq'})^2}{M_p^2}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (28{\cdot}11)$$

and

$$\mu_2(D_4^2) = \frac{4}{P^2} S_p\left[\frac{(\bar{m}_{pq} - \bar{m}_{pq'})^2}{M_p^2} \cdot \frac{\bar{\sigma}_p^2}{M_p^2}\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)\right] + \frac{2}{P^2} S_p\left[\frac{\bar{\sigma}_p^4}{M_p^4}\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)^2\right] .. \quad (28{\cdot}2).$$

When the size of the samples are constant, we may write $\frac{2}{\bar{n}_q} = \left(\frac{1}{n_{pq}} + \frac{1}{n_{pq'}}\right)$, and also writing

$$V_2 \equiv \frac{1}{P} S_p\left[\left(\frac{\bar{\sigma}_p^2}{M_p^2}\right)\right], \text{ and } V_4 \equiv \frac{1}{P} S_p\left[\left(\frac{\bar{\sigma}_p^4}{M_p^4}\right)\right] \quad .. \quad (28{\cdot}3)$$

we have

$$D_4^2=\frac{1}{P}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{M_p^2}\right]-\frac{8}{\bar{n}_q}\cdot V_2 \quad .. \quad .. \quad .. \quad (28\cdot4)$$

and

$$\mu_2(D_4^2)=\frac{8}{P\cdot\bar{n}_q}S_p\left[\frac{(\bar{m}_{pq}-\bar{m}_{pq}')^2}{M_p^2}\cdot\left(\frac{\bar{\sigma}_p^2}{M_p^2}\right)\right]$$
$$+\frac{8}{P\cdot\bar{n}_q^2}\cdot V_4 \quad .. \quad (28\cdot5).$$

Similar expressions can be obtained for a fifth coefficient D_5^2 by using the "familial" mean (m_p) instead of the inter-class mean (M_p).

VII. The Principle of Equipartition of Variance and a Coefficient of Familial Differentiation

27. There exist certain algebraic relations between $\bar{\sigma}_p^2$, s_p^2 and Σ_p^2 which are of considerable interest in connection with the question of the choice of a suitable value for k_p^2.

We start with the algebraic identity:

$$S_qS_t[(x_{pqt}-m_p)^2]=S_qS_t[(x_{pqt}-m_{pq})^2]+$$
$$S_q[n_{pq}(m_{pq}-M_p)^2] \quad .. \quad .. \quad (29\cdot0).$$

Using equations (2·8) and (2·4) we have from the above equation

$$n_p\cdot\Sigma_p^2=S_q[(n_{pq}\cdot\sigma_{pq}^2)]+S_q[n_{pq}(m_{pq}-M_p)^2] \quad .. \quad .. \quad (29\cdot1).$$

If the size of the sample is kept invariable for the same character for all the samples, *i.e.*, n_{pq} is constant for all values of q, we get

$$n_p\cdot\Sigma_p^2=n_{pq}\cdot S_q[(\sigma_{pq}^2)]+n_{pq}\cdot S_q[(m_{pq}-M_p)^2] \quad .. \quad .. \quad (29\cdot2).$$

Using (2.9) and (2.6) we obtain

$$n_p\Sigma_p^2=n_{pq}\cdot(N_p\bar{\sigma}_p^2)+n_{pq}\cdot(N_ps_p^2) \quad .. \quad .. \quad (29\cdot3).$$

But when n_{pq} is constant for all values of q, we have $n_p=N_p\cdot n_{pq}$, by equation (2·2). We therefore get finally

$$\Sigma_p^2=\bar{\sigma}_p^2+s_p^2 \quad .. \quad .. \quad .. \quad .. \quad (29\cdot4).$$

The total or familial variance is made up of the average variance within the group (when the size of the group is kept constant) together with the variance for variation from group to group.

28. Equation (29·4) is very suggestive. Consider a population which has become differentiated in course of time into a large number of different groups. It leads us to enunciate a proposition that, when the variation has proceeded in an absolutely random manner (and for a sufficiently long time), the total variance within the population would tend to become equally distributed between the different modes of variation.

We may refer to this proposition as the *principle of equipartition of variance*. When the variance within any particular group reaches a certain limiting value conditions would become unstable, and the group would tend to break up into two or more sub-groups. On the other hand if the variation within a group becomes too restricted, the group itself would tend to disappear or become absorbed by other groups. For absolutely random variation therefore we may expect that

$$\Sigma_p^2 = 2\bar{\sigma}_p^2 = 2s_p^2 \quad .. \quad .. \quad .. \quad (30{\cdot}0).$$

29. The ratio of the inter-group[1] variance (s_p^2) to the average intra-group variance ($\bar{\sigma}_p^2$) would thus furnish a convenient coefficient for the measurement of the differentiation within any given collection (or family) of groups. We may call such a quantity a coefficient of familial differentiation and define it by

$$f^2 = \frac{1}{P} S\left[\left(\frac{s_p^2}{\sigma_p^2}\right)\left(\frac{1}{1+\eta_p}\right)\right] \quad .. \quad .. \quad (31{\cdot}0)$$

where

$$v_p^2 = \frac{\Sigma s_p^2}{\bar{s}_p^2} = \frac{1}{2 \cdot N_p}, \quad w_p^2 = \frac{\Sigma \sigma_p^2}{\bar{\sigma}_p^2} = \frac{1}{2 \cdot n_p} \quad .. \quad (31{\cdot}01)$$

and $$(1+\eta_p) = (1+v_p^2)(1+3w_p^2+15w_p^4+105w_p^6 + 945w_p^8 + 10,395w_p^{10}) \quad .. \quad .. \quad .. \quad (31{\cdot}02).$$

Here N_p and n_p are the effective numbers of observation on which the estimates of the two variances s_p^2 and $\bar{\sigma}_p^2$ are respectively based.

For any given collection of groups, N_p the number of groups will usually be considerably smaller than n_p which is a number of the order of the total number of individuals in the whole collection, and $\frac{1}{2 \cdot n_p}$ may therefore be neglected in comparison with $\frac{1}{2N_p}$.

---

[1] It would be better to call it the co-group variance, so that it may be clearly distinguished from the intra-group variance.

Thus when $w_p^2=0$ approximately, we have

$$f^2=\frac{1}{P}S_p\left[\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)\left(\frac{1}{1+v_p^2}\right)\right] \quad .. \quad .. \quad (31\cdot1)$$

with mean value

$$\bar{f}^2=\frac{1}{P}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)\right] \quad .. \quad .. \quad .. \quad (31\cdot11)$$

where $\bar{s}_p^2$ and $\bar{\sigma}_p^2$ are mean values of s_p^2 and σ_p^2 respectively. Also using equations (10·6)–(10·9) and (11·1) –(11·4),

$$\mu_2(f^2)=\frac{2}{P^2}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^2\cdot\left\{v_p^2(2-3v_p^2-4v_p^4-5v_p^6+6v_p^8)\right\}\right] \quad .. \quad .. \quad .. \quad (31\cdot2)$$

$$\mu_3(f^2)=\frac{8}{P^3}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^3\cdot\left\{v_p^4(3-8v_p^2+15v_p^4-24v_p^6)\right\}\right] \quad .. \quad .. \quad (31\cdot3)$$

$$\mu_4(f^2)=\frac{12}{P^4}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^4\cdot\left\{v_p^4(4+4v_p^2-35v_p^4+100\,.\,v_p^6)\right\}\right]$$

$$+\frac{24}{P^4}S_pS_r\left[\left\{\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)\cdot v_p^2(2-3v_p^2+4v_p^4+5v_p^6+6v_p^8)\right\}\left\{\left(\frac{\bar{s}_r^2}{\bar{\sigma}_p^2}\right)\cdot v_p^2(2-3v_r^2+4v_r^4-5v_r^6+6v_r^8)\right\}\right] \quad .. \quad (31\cdot4).$$

When $N_p=N$ is constant for all values of p, we may write the above equations in the following form:

$$f^2=\left(\frac{1}{1+v^2}\right)\cdot\frac{1}{P}S\left[\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)\right]=\frac{\gamma_1}{P}S_p\left[\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)\right] \quad .. \quad .. \quad (31\cdot5)$$

$$\text{with} \quad \bar{f}^2=\frac{1}{P}S_p\left[\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (31\cdot51)$$

$$\mu_2(f^2)=\frac{2\gamma_2}{N\,.\,P^2}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^2\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (31\cdot6)$$

$$\mu_3(f^2)=\frac{6\,.\,\gamma_3}{N\,.\,P_3}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^3\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (31\cdot7)$$

$$\mu_4(f^2)=\frac{12\,.\,\gamma_4}{N\,.\,P^4}S_p\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)^4\right]+\frac{24\gamma_2^2}{N^2\,.\,P^4}S_pS_r\left[\left(\frac{\bar{s}_p^2}{\bar{\sigma}_p^2}\right)\left(\frac{\bar{s}_r^2}{\bar{\sigma}_r^2}\right)\right] \quad (31\cdot8)$$

where γ_1, γ_2, γ_3, and γ_4 are purely numerical factors given by

$$\left.\begin{aligned}\gamma_1&=1-\frac{1}{2N}+\frac{1}{4N^2}-\frac{1}{8N^3}+\frac{1}{16N^4}\\ \gamma_2&=1-\frac{3}{2N}+\frac{1}{2N^2}-\frac{5}{16N^3}+\frac{3}{16N^4}\\ \gamma_3&=1-\frac{4}{3N}+\frac{5}{4N^2}-\frac{1}{N^3}\\ \gamma_4&=1+\frac{2}{N}-\frac{85}{16N^2}+\frac{25}{8N^3}\end{aligned}\right\} \quad .. \quad .. \quad (31{\cdot}9).$$

The numerical factors, γ_1, γ_2, γ_3, and γ_4 approach the limiting value 1 as N increases. I give below actual values of the coefficients for a few selected values of N.

$N=$	10	20	50	100	∞
γ_1	0·952	0·976	0·990	0·995	1·000
γ_2	0·855	0·926	0·970	0·985	1·000
γ_3	0·878	0·934	0·973	0·987	1·000
γ_4	1·181	1·095	1·039	1·020	1·000

When $\left(\frac{\bar{s}_p^2}{\sigma_p^2}\right)=1$ for all values of p, the above formulæ reduce to

$$\left.\begin{aligned}&\bar{f}^2=1,\ \mu_2(f^2)=\frac{2\gamma_2}{N\,.\,P},\ \mu_3(f^2)=\frac{6\gamma_3}{N^2P^2}\\ &\mu_4(f^2)=\frac{12\gamma_4}{N^2P^3}+\frac{12\gamma_2{}^2(P-1)}{N^2P^3}\end{aligned}\right\} \quad .. \quad .. \quad (31{\cdot}91).$$

If in addition N is large,

$$\beta_1=\frac{9}{2NP}\left(1+\frac{11}{6N}\right),\ \beta_2=3+\frac{15}{PN} \quad .. \quad .. \quad (31{\cdot}92).$$

30. The usefulness of the coefficient of familial differentiation (f^2) does not of course depend on the validity or otherwise of what I have called the principle of equipartition of variance. For it is very simply connected with D^2, the first coefficient of divergence defined by equation (14·0).

We may write

$$(m_{pq}-m_{pq}')=(m_{pq}-M_p)+(M_p-m_{pq}') \quad .. \quad (32{\cdot}1).$$

It is easily seen that squaring equation (32·1), and summing for all possible comparisons within the given collection (or family) of groups we get

$$S[(m_{pq}-m_{pq}')^2]=(q-1)S[(m_{pq}-M_p)^2]=(q-1)\,.\,q\,.\,s_p^2\,.\,. \quad (32\cdot2)$$

where $q\,.\,s_q^2=S[(m_{pq}-M_p)^2]$ gives the co-group variance for the pth character, q being the total number of groups available. For any assigned character the mean value of $(m_{pq}-m_{pq}')^2$ for all possible comparisons is thus $2s_p^2$, since the total number of comparisons possible$=q(q-1)/2$. Writing the mean value of D^2 for all possible comparisons ($q(q-1)/2$ in number) as D_m^2, we have therefore

$$D_m^2=\frac{1}{P}\,S\left[\left(\frac{2s_p^2}{\bar{\sigma}_p^2}\right)\right]-\frac{2}{\bar{n}}=\left[2(1+v^2)\,.\,f^2\right]-\frac{2}{\bar{n}} \quad .. \quad (32\cdot3)$$

where $v^2=\dfrac{1}{2N}$ approximately, and $\bar{n}$ is the harmonic mean of all the different values of n_{pq}, n_{pq}'.

When $\bar{n}$ and N are both large, *i.e.*, $\dfrac{2}{\bar{n}}$ and $\dfrac{1}{2N}$ are both negligibly small, we shall have $D_m^2=2f^2$. We thus find that (excepting for a small correcting factor) the coefficient of familial differentiation f^2 is numerically equal to half the *average* value of the group divergence, the average being taken for all possible comparisons within the given collection of groups.

For an absolutely random collection, if we assume that the principle of equi-partition of variance is true, we should have

$$s_p^2=\bar{\sigma}_p^2=\tfrac{1}{2}\Sigma_p^2 \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (32\cdot4).$$

In this case the two measures of divergence D^2 and D_2^2 become equal, while

$$D_3^2=\tfrac{1}{2}D_2^2=\tfrac{1}{2}D^2 \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad .. \quad (32\cdot5).$$

VIII. A Coefficient of Divergence in Variability

31. We may also easily construct measures of divergence in variability. Let us choose $x_{pq}=\sigma_{pq}$, $x_{pq}'=\sigma_{pq}'$, and $k_p^2=\bar{\sigma}_p^2$, where $\bar{\sigma}_p^2$ is a reliable *constant* value of the variance in the pth character which does not fluctuate from sample to sample. Then as a first approximation we may substitute

$$\Sigma x_{pq}^2=\frac{\bar{\sigma}_p^2}{2n_{pq}}\,,\quad \Sigma x_{pq}'^2=\frac{\bar{\sigma}_p^2}{2n_{pq}'} \quad .. \quad .. \quad .. \quad .. \quad (33\cdot01)$$

$$v_p^2=\frac{\bar{\sigma}_p^2}{(\bar{\sigma}_{pq}-\bar{\sigma}_{pq}')^2}\left(\frac{1}{2n_{pq}}+\frac{1}{2n_{pq}'}\right),\ w_p^2=0 \quad .. \quad (33\cdot02)$$

where n_{pq}, n_{pq}' are the size of the two samples.

Using equation (13) we define a measure of divergence in variability by

$$F^2=\frac{2}{P}S_p\left[\frac{(\sigma_{pq}-\sigma_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{2}{P}S_p\left[\left(\frac{1}{2n_{pq}}+\frac{1}{2n_{pq}'}\right)\right]. \quad (33{\cdot}0)$$

$$=\frac{2}{P}S_p\left[\frac{(\sigma_{pq}-\sigma_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{2}{\bar{n}_q} \quad .. \quad .. \quad .. \quad .. \quad .. \quad (33{\cdot}01)$$

when the size of the samples is constant for all characters, or a mean value $\bar{n}_q$ as defined by equation (16·7) may be used without appreciable errors. The mean value of F^2 is easily found to be

$$\bar{F}^2=\frac{1}{P}S_p\left[\frac{(\bar{\sigma}_{pq}-\bar{\sigma}_{pq}')^2}{\bar{\sigma}_p^2}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (33{\cdot}1)$$

$$\mu_2(F^2)=\frac{16}{P^2}S_p\left[\frac{(\bar{\sigma}_{pq}-\bar{\sigma}_{pq}')^2}{\bar{\sigma}_p^2}\left(\frac{1}{2n_{pq}}+\frac{1}{2n'_{pq}}\right)\right]$$

$$+\frac{8}{P^2}S_p\left[\left(\frac{1}{2n_{pq}}+\frac{1}{2n_{pq}'}\right)^2\right] \quad .. \quad (33{\cdot}2).$$

We may use equation (16·7), and write

$$\frac{1}{\bar{n}_q}=\frac{1}{P}S_p\left[\left(\frac{1}{2n_{pq}}+\frac{1}{2n_{pq}'}\right)\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (16{\cdot}71)$$

when the size of the samples is constant for all character, or when the fluctuation in the size of the sample can be neglected.

We then obtain

$$\mu_2(F^2)=\frac{8}{P\,.\,\bar{n}_q}\left[2(\bar{F}^2)+\frac{1}{\bar{n}_q}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (33{\cdot}21)$$

$$\mu_3(F^2)=\frac{64}{P^2\,.\,\bar{n}_q^2}\left[3(\bar{F}^2)+\frac{1}{\bar{n}_q}\right] \quad .. \quad .. \quad .. \quad .. \quad .. \quad (33{\cdot}3)$$

$$\mu_4(F^2)=\frac{768}{P^2\,.\,\bar{n}_q^3}\left[(\bar{F}^2)^2\right]+\frac{192(P+4)}{P^3\,.\,\bar{n}_q^3}\left[4(\bar{F}^2)+\frac{1}{\bar{n}_q}\right]. \quad (33{\cdot}4).$$

When the two samples are drawn from the same group or population, or when there is no significant difference in variability we have $(\bar{\sigma}_{pq}-\bar{\sigma}_{pq}')=0$ for all values of p, and we get

$$(\bar{F}^2)_0=0\pm{\cdot}67449\,.\,\frac{2}{\bar{n}_q}\sqrt{\frac{\overline{2}}{P}} \quad .. \quad .. \quad (33{\cdot}5)$$

a formula which is analogous to equation (16·8).

If the variances for different samples are widely different, and it is not considered desirable to use $k_p^2=\bar{\sigma}_p^2$, we may still have recourse to the present method, and develop approprite

formulæ by substituting $k_p^2=s_p^2$ in equations (13.0) and (12·1)–(12·4).

32. Coefficients of divergence in Skewness, and Kurtosis may also be constructed with the help of equation (13·0). For example for β_2, we may use $x_{pq}=\beta_2(p, q)$, $x_{pq}'=\beta_2'(p, q')$, etc.

In the case of β_1 and β_2 (or other β-constants) a simpler alternative is open to us. It will be remembered that β_1, β_2 and the other β-constants are pure numbers, so that the difficulty due to non-homogeneity of dimensions, discussed in paragraph 7, does not exist in their case. We may therefore use a coefficient of a simpler form, by putting $k_p^2=1$, and using

$$\frac{1}{P}S_p\left[\left\{\beta_2(p, q)-\beta_2'(p, q')\right\}^2\right]$$

with a small correcting term to allow for the bias introduced by the finite size of the samples.

I may also point out that the need for these coefficients will usually arise only when both C^2 and E^2(defined by equation (5·1) have failed to reveal the existance of divergence. In such cases it will also be usually sufficient to employ ordinary tests of statistical divergence between the corresponding β-constants for the two samples for each character separately. It must be remembered however that divergence in β-constants can be tested (or measured) only when the size of the samples is very large.

IX. Conclusion

33. It will be useful to have at this stage a brief resume of the important formulæ.

A convenient measure of divergence in means is given by

$$D^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{1}{P}S_p\left[\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad .. \quad (14·0)$$

and subsidiary equations (14·1)–(14·4). Modified values under restricted conditions are given in equations (16·0)–(16·8), while more general values are given in equations (27·1)–(27·8). Results of experimental sampling discussed in Section 5 are in satisfactory agreement with the theory.

A second measure of divergence is furnished by

$$D_2^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{s_p^2}\right]-\frac{1}{P}S_p\left[\frac{\bar{\sigma}_p^2}{s_p^2}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad .. \quad (26·0)$$

and equations (26·1)–(26·5).

An exactly analogous coefficient may be constructed by using the familial variance Σ_p^2:—

$$D_3^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')^2}{\Sigma_p^2}\right]-\frac{1}{P}S_p\left[\frac{\bar{\sigma}_p^2}{\Sigma_p^2}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad .. \quad (35·0).$$

In certain ways D_3^2 would be an extremely convenient coefficient. Unfortunately, owing to lack of sufficient data in anthropology, it is not possible to obtain reliable values of Σ_p^2 for the whole human species. Neither is it possible, for the same reason, to obtain reliable values of s_p^2 for the human species. Fairly reliable values of the intra-group variance $\bar{\sigma}_p^2$ may however be calculated in many cases, and the coefficient D^2 may therefore be used without difficulty.

I have given certain reasons for believing that D^2 would under certain conditions give practically the same results as D_2^2 or D_3^2. When variation may be supposed to have taken place in an absolutely random manner within a given collection of groups, (say within the human species), a plausible hypothesis is that the total variance would tend to be distributed equally between the variation within the groups and the variation from group to group. In this case

$$\Sigma_p^2 = 2\bar{\sigma}_p^2 = 2s_p^2 \quad .. \quad .. \quad .. \quad (35{\cdot}1)$$

so that

$$D_3^2 = \tfrac{1}{2} D^2 = \tfrac{1}{2} D_2^2 \quad .. \quad .. \quad .. \quad (35{\cdot}2).$$

When sufficient data become available it will be possible to test the above theory.

34. In case however no such simple relation (as predicted above) is found in future to subsist between D^2, D_2^2, and D_3^2, the choice between these coefficients would have to be made by reference to the respective results obtained by their use.

The great simplicity of the equations for D^2 will, however, remain an important point in its favour; and other things being equal or nearly equal, this will be a sufficient reason for its general adoption.

Another convenient property of D^2 is that it may be easily converted into the Pearsonian Coefficient of Racial Likeness (C^2) by multiplication with suitable numerical factors. When n_{pq}, n_{pq}' are constant for all characters or the fluctuation in the size of the sample can be neglected it will be noticed that

$$C^2 = \left(\frac{n_q \cdot n_q'}{n_q + n_q'}\right) . D^2 = \frac{2}{\bar{n}_q} . D^2 \quad .. \quad .. \quad (36{\cdot}1).$$

When n_{pq}, n_{pq}' are not constant, and the size of the samples cannot be neglected, we still have

$$C^2 = \frac{1}{\bar{P}} S_p \left[\frac{(m_{pq} - m_{pq}')^2}{\bar{\sigma}_p^2}\left(\frac{1}{n_{pq}} + \frac{1}{n_{pq}'}\right)\right] - 1 \quad \ldots \quad (36{\cdot}2)$$

in which the terms $(m_{pq} - m_{pq}')^2/\bar{\sigma}_p^2$ will have already been calculated for the computation of D^2.

The use of D_2^2 would appear to be indicated where a close study of the differentiation *within* a given family is required. It has the great advantage that, on the average of all possible

comparisons within a collection, it gives the same "weight" to all characters, i.e., does not discriminate against any particular set of characters. A consequential disadvantage is of course that the average value of D_2^2 (for all possible comparisons within a collection) remains identically same for every collection, so that a comparison of coefficients from different collections (or families) may become extremely misleading.

A different type of the coefficient of divergence in means is given by

$$D_4^2=\frac{1}{P}S_p\left[\frac{(m_{pq}-m_{pq}')}{M_p^2}\right]-\frac{1}{P}S_p\left[\frac{\bar{\sigma}_p^2}{M_p^2}\left(\frac{1}{n_{pq}}+\frac{1}{n_{pq}'}\right)\right] \quad .. \quad (28{\cdot}1)$$

and equations (28·2) – (28·5).

35. We have also proposed to use the ratio $\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)$ for measuring the amount of differentiation existing within a given collection of groups, and we have defined a coefficient of familial differentiation by

$$f^2=\frac{1}{P}S_p\left[\left(\frac{s_p^2}{\bar{\sigma}_p^2}\right)\left(1-\frac{1}{2N_p}+\frac{1}{4N_p^2}-\frac{1}{8N_p^3}\right)\right] \quad .. \quad .. \quad (31{\cdot}1)$$

and equations (31·0) – (31·8), where N_p is the number of groups included in the comparison. This coefficient is very simply connected with the average value D_m^2 of the first coefficient of divergence (the average being taken for all possible comparisons within the given collection of groups).

$$D_m^2=2(1+\gamma^2)\,.f^2-\frac{2}{\bar{n}_q} \quad .. \quad .. \quad .. \quad (32{\cdot}8)$$

$(1+\gamma^2)$ being a numerical factor which approaches the limiting value 1 as N increases.

38. Apart from the Pearsonian Coefficient of Racial Likeness (which furnishes the standard test for the detection of divergence in means), we have obtained several new tests of divergence. The most important of which is a coefficient for detecting divergence in variabilities, which may be used in practice without difficulty.

$$E^2=\frac{2}{P}S_p\left[\left(\frac{n_{pq}\,.\,n_{pq}'}{n_{pq}+n_{pq}'}\right)\frac{(\sigma_{pq}-\sigma_{pq}')^2}{\bar{\sigma}_p^2}\right]-1 \quad .. \quad (5{\cdot}1).$$

A convenient *measure* of divergence in variabilities is given by

$$F^2=\frac{2}{P}S_p\left[\frac{(\sigma_{pq}-\sigma_{pq}')^2}{\bar{\sigma}_p^2}\right]-\frac{2}{P}S_p\left[\left(\frac{1}{2n_{pq}}+\frac{1}{2n_{pq}'}\right)\right] \quad .. \quad (33{\cdot}0)$$

and subsidiary equations (33·1) – (33·4).

It will be noticed that when n_{pq} and n_{pq}' are constant for all values of p, or the fluctuation in the size of the sample is negligibly small,

$$E^2=\left(\frac{n_q \cdot n_q'}{n_q+n_q'}\right).F^2=\frac{2}{\bar{n}_q}.F^2 \quad .. \quad .. \quad (38{\cdot}0)$$

a result which can be compared with that given in equation (36·1).

36. I wish to emphasize here the distinction between tests and measures of divergence. It is true that the Pearsonian Coefficient of Racial Likeness (which is properly speaking a test of divergence) has been extensively used with considerable success as a measure of divergence in craniometry. This point will be discussed later, but a little consideration will show that such use can be considered legitimate only under restricted conditions.

Consider two samples drawn from the *same* group or population. In this case we must have

$$C^2=0\pm{\cdot}67449\sqrt{\frac{\overline{2}}{P}} \quad .. \quad .. \quad .. \quad (39{\cdot}0).$$

In using the above equation to detect the existence of divergence we adopt the following procedure:—

(i) We assume that the two samples under consideration are drawn from the *same* group or population; *i.e.* $(\bar{m}_{pq}-\bar{m}_{pq}')=0$ for all characters. ..... (Hypothesis (A).

(ii) Then by comparing the observed value of C^2 with equation (39·0) we now determine the probability of hypothesis (A) being true.

If C^2 is not significantly different from zero, we are in a position to assert that, judged by the given data, the two groups (from which the two samples were drawn) are probably not different. On the other hand if C^2 is significantly greater than zero, we feel justified in asserting that the two groups are differentiated from each other.

The point to be noted here is that the magnitude of C^2 determines the degree of certainty with which the *existence* of divergence can be asserted, but does not necessarily supply any information regarding the *magnitude* of such divergence.

So long as the samples are drawn from the same group or population, C^2 will be approximately equal to zero whatever be the value of n or n'. When the two samples are however drawn from two different groups or populations, $(\bar{m}_{pq}-\bar{m}_{pq}')$ would not vanish generally, and $\bar{D}^2$ would attain some constant finite value for the same two differing groups. The observed values of C^2 would in such cases depend upon both

(i) the magnitude of $\bar{D}^2$, as well as on

(ii) the factor $\left(\frac{n \cdot n'}{n+n'}\right)$ determined by the size of the samples.

Provided n, n' are fairly large, observed values of D^2 would differ from the mean value $(\bar{D}^2)$ by quantities of the order of errors of random sampling, *i.e.* by quantities which will be in usual practice negligibly small; so that observed values of D^2 would not differ significantly from $(\bar{D}^2)$. Thus the first factor D^2 would remain sensibly constant for samples drawn from the same two differing populations. The factor $\left(\frac{n \cdot n'}{n+n'}\right)$ would however vary directly with the size of the samples, so that observed values of C^2 also would vary with the size of samples, and would not remain sensibly constant for the same two differing populations.

If the size of the samples n, n' are very large we may easily obtain very large values of C^2 even when the samples are drawn from two groups which are closely associated. On the other hand when n, n' are small C^2 may assume very small values even for widely divergent populations.

This difficulty (and the need for making allowances for the size of the samples) was recognised long ago by G. H. Morant (7, p. 12) who wrote:—

"Given two random samples each of ten individuals drawn from the same homogeneous population, the Coefficient of Racial Likeness .... deduced from the mean character of the two samples will not differ significantly from zero, and if two samples each of a hundred individuals are drawn from the same population then their coefficients will also be of the same order.[1] But if two random samples each of ten individuals are drawn from two different populations and then two samples each of a hundred individuals are drawn from the same differing populations it will be found that the coefficient between the first pair will be very distinctly less than that between the two samples of hundred individuals each. The difference in this case is merely an expression of the rather obvious fact that it is more probable that the small samples were in reality drawn from the same population than that the larger samples were. It is for this reason that the coefficients of Racial Likeness may not be compared directly by estimating differences in terms of probable errors only as may be done when dealing with the majority of statistical constants in use. Reference has to be made constantly to the number of crania in the several racial series used.

---

[1] *i.e.*, will be of the order of zero (P.C.M.).

..... In practice direct comparison may be made between the numerical values of the coefficients ..... in the cases when":—

(i) "All the .... means are based on the same or approximately the same number of crania; and"

(ii) "When one series of racial means is compared with a number of others, the latter being based on the same or approximately the same numbers of crania which may differ from the first series."

It will be seen that Dr. Morant's two conditions may be combined into the single statement that the factor $\left(\frac{n \cdot n'}{n+n'}\right)$ must remain sensibly constant during the same series of comparisons. Enforcement of this restriction would therefore (apart from errors of random sampling) inevitably throw the comparison to the factor D^2.

It is clear from the above discussion that the use of C^2 as a measure of divergence would be strictly possible only under either of the following two conditions:—

(*a*) when the samples are drawn from the same population, *i.e.*, when C^2 is sensibly zero, or

(*b*) when the size of the samples, and hence the factor $\left(\frac{n \cdot n'}{n+n'}\right)$ remains constant for all samples.

37. Prof. Pearson (12, 105-117) has shown however in a review of about 750 computed values of C^2 that in actual practice the Coefficient of Racial Likeness has been found to be an extremely useful tool in craniometric researches. For purposes of comparison I therefore obtained by direct calculation approximate values of D^2 corresponding to nearly every one of the 750 values of C^2 reviewed by Prof. Pearson. My results will be fully discussed in Part II of the present paper, but I may anticipate a little and state here, that I believe I have succeeded in tracing the empirical success of the use of C^2 as a measure of divergence in craniometric work in most cases to either or both of the conditions explained above. A very large number of the coefficients (reviewed by Prof. Pearson) referred to closely associated groups for which both C^2 and D^2 gave coefficients of low magnitudes. Further, owing to paucity of material the number of skulls in each sample was also usually small, so that the size factor $\left(\frac{n \cdot n'}{n+n'}\right)$ did not fluctuate very widely. In fact I could detect only a comparatively small number of coefficients for which C^2 and D^2 gave significantly different results.

Conditions are however widely different for measurements on the living; the size of the sample is more variable, and much

larger samples are often met with in practice, so that the influence of the factor $\left(\frac{n \cdot n'}{n+n'}\right)$ is not negligible.

I felt this difficulty several years ago, and in order to avoid it had used $\frac{1}{P} S_p \left[\frac{(m_{pq} - m_{pq}')^2}{\bar{\sigma}_p^2}\right]$ as a measure of divergence in two anthropometric papers, one on the "Chinese Head" (6) and the other on "Race Mixture in Bengal" (5). The results obtained were I believe fairly encouraging.

In the present paper I have obtained coefficients which are theoretically[1] preferable to the one used by me previously, and I have also investigated their statistical distributions. I have made an empirical study of five of the coefficients (D^2, D_2^2, D_4^2, E^2, and F^2, defined by equations (14·0), (26·0), (28·1), (5·1) and (33·0) respectively) using a long series of Swedish measurements on the living (4). The results will be given in a sequel to the present paper,[2] but I may mention in anticipation that they support the use of D^2 for comparative purposes.

Addendum

In June 1927, I showed a first draft of the present paper to Prof. Karl Pearson, and discussed with him the difficulties connected with the fluctuating size of samples $\left(\frac{n \cdot n'}{n+n'}\right)$. At that time he was unable to accept my views, and he pointed out certain theoretical objections to my results. I then worked out the mathematical portion with greater rigour, and communicated the present paper to the Indian Science Congress in December 1928. About the same time Prof. Pearson himself proposed (12*a*)[3] making allowances for the size factor $\left(\frac{n \cdot n'}{n+n'}\right)$ by reducing all coefficients of Racial Likeness to a standard population. When the size of the sample is constant for all characters, the result of such reduction would be to make the

---

[1] I would point out that the theoretical limitations given by the set of assumptions (A-1)-(A-7) under which the present formulæ have been worked out are practically the same as those subsisting for the Pearsonian Coefficient C^2. These restrictions have been fully discussed by Prof. Pearson (12). The most important of the restrictions which requires further consideration is the neglecting of the correlation between different characters.

[2] The anthropological portion of the work on the Swedish material has been published in the *Biometrika*, Vol. XXII, 1930, 94-108 ("A Statistical Study of certain Anthropometric Measurements from Sweden").

[3] This part of the *Biometrika* reached me in Calcutta in March, 1929.

reduced values of C^2 (the Pearsonian Coefficient of Racial Likeness) strictly proportional to D^2 (the Coefficient of Divergence described in the present paper). Even when the size of the sample is not constant for all characters the reduced C^2 would still be approximately proportional to D^2, so that in actual practice both coefficients would usually yield very nearly the same results. There is, however, one definite advantage in favour of D^2; its probable error can be calculated without difficulty, and hence values of D^2 can be compared directly.

References

1. Elderton, W. Palin .. "Frequency Curves and Correlations," 2nd Edition, (C. and E. Layton, 1927).
2. Fisher, R. A. .. "Statistical Methods for Research Workers," 2nd. Edition, (Oliver and Boyd, 1928).
3. ,, ,, .. "Mathematical Foundations of Theoretical Statistics." Phil. Trans., Vol. XII, CCXXII, 309-368.
4. Lundborg, H. and Linders, F. J. .. "Racial Characters of the Swedish Nation." (Swedish Institute for Race Biology, Upsala, 1926).
5. Mahalanobis, P. C. .. "Analysis of Race Mixture in Bengal." (Jour. and Proceedings, Asiatic Soc. of Bengal, Vol. XXIII, 1927, No. 3).
6. ,, ,, .. "A Statistical Study of the Chinese Head.' (Man In India) Vol. VIII, April and Sept. 1928.

6(*a*). ,, ,, .. "A Statistical Study of Certain Anthropometric Measurements from Sweden." Biom., Vol. XXII, Parts I and II, 1930, 94–108.

7. Morant, G. H. .. "A study of Certain Oriental Series of Crania," etc. (Biom., Vol. XVI, 1924, 1–105).
8. Neyman, J. and Pearson, E. S. .. "On the use and Interpretation of certain test Criteria for purposes of Statistical Inference." Biometrika, Vol. XXA, 1928, 175–240.
9. Pearon, E. S .. "A Further Note on the Distribution of Range in Samples taken from Normal Population." Biometrika, Vol. XVIII, 1926, 173–194.
10. Pearson, Karl (Edited by) .. .. "Tables For Statisticians and Biometricians." (Cambridge University Press, 1914).
11. Pearson, Karl .. "On the Probability that two Independent Distributions of Frequency are really samples from the same Population." Biom., Vol. VIII, 250–254.
12. ,, ,, .. "On the Coefficients of Racial Likeness." Biometrika, Vol. XIII, 1926, 105–117.

12(*a*). Pearson, Karl .. "Note on Standardisation of Method of using the Coefficient of Racial Likeness." Biom., Vol. XXB, 1928, 376–378.

13. Tchouproff, Al. A. .. "On the Mathematical Expectation of the Moments of Frequency Distributions." Part II, Biom., Vol. XIII, 1921, 283–295.

14. Tildesley, M. L. .. "A First Study of the Burmese Skull." Biom., Vol. XIII, 1921, 247–251.

15. Tippett, L. H. C. .. "Random Sampling Numbers." Tracts for Computors No. XV, (Cambridge Univ. Press, 1927).

16. ,, ,, .. "On the Extreme Individuals and the Range of Samples taken from a Normal Population." Biometrika, Vol. XVII, 1925, 364–383.

17. Yasukawa, Kazutaro .. "On the Probable Error of the Mode of Skew Frequency Distributions." Biometrika, Vol. XVIII, 1926, 262-292.