CORRELATIONS OF SUMS OR DIFFERENCES.

By C. SPEARMAN.

(*From the Psychological Laboratory, University College, University of London.*)

After calculating the correlations between several series of values, it frequently happens that we want the correlations given by some of the series added together; or, what comes to the same thing, we want the correlations of the average of some of the series.

Suppose, for instance, that we have tested a number of children on two separate occasions. We are bound to work out the two sets of results separately, in order to ascertain how far they differ from one another, and whether the differences indicate a change of experimental conditions or may reasonably be ascribed to mere 'chance.' But having done this, we next generally desire the result of taking the mean of both occasions. Or again, suppose that we have measured the correlations of the accuracy and also of the speed of any performance; it is almost always interesting to regard the performance as a whole, allotting marks partly for speed and partly for accuracy. Or once more, supposing that we have found out the correlations between a number of experimental tests and position in school, we may wish to learn how far the school position correlates with all the tests pooled together.

Scarcely less important than the correlations between sums are those between differences. When, for instance, children have been tested twice, our chief interest might be in their improvement; we wish to get the correlations of this improvement, that is, of the remainder obtained by subtracting the first result from the second.

But the calculations involved in obtaining the correlations for the sums or the differences are generally laborious. Before we can start working out the further coefficients required, we have to make the necessary additions or subtractions of the original values. And before commencing even this operation, we are obliged to reduce these values to suitable proportions to one another; if, as is usually the case, we

desire all the series compounded to have an equal influence on any ensuing correlation, we must multiply them all by such factors as will make their 'standard deviations[1]' equal.

There are many problems, too, where our long work would not even be successful in the end. Suppose, in the instance above, that we were not content to correlate the school position with a pool in which all the tests stood on an equal footing, but aspired to find the proportions in which these tests should be combined in order to make the correlation with school work as large as possible. Such a problem would rarely admit of satisfactory direct attack, however much labour we were willing to expend.

The desired calculation may be not only unfeasible in ordinary practice but essentially impossible. In fact, one may say that this almost invariably happens; for nearly every value attainable by us is more or less vitiated by being only a representative *sample* of the entire class with which we really wish to deal. We should like to obtain and pool such values in infinite number.

Take, for instance, all correlations between mental performances; the power exhibited by a person on any occasion, whether in discriminating pitch or in memorising poetry, in discursive reasoning or in aesthetic combination, is no more than a sample of his general power for the kind of act in question; another sample might well yield a different result. Or the required correlation may be between series of frequencies; for instance, we may want to know how far the tendency in children to be 'naughty' in bad weather is a sign of neurotic disposition; the frequencies actually observed are necessarily infinitely few as compared with the whole class under consideration. The same may be said of correlations between series of correlations. In all cases alike, the desired results of universal validity would have to be based on an infinite number of samples or cases. Actually to obtain and pool these is, of course, out of the question; but at any rate, it is conceivably possible and eminently desirable to calculate the most probable value of such a pool.

The aim of this paper, then, is to save labour and to overcome obstacles to research by expressing the required correlation between sums (or differences) as a simple function of the correlations between the elements combined into these sums (or differences). Let the two groups of elements to be combined be denoted by $a_1, a_2, \ldots a_p$ and

[1] Using this expression as a short equivalent for 'root mean square deviation.'

$b_1, b_2, \ldots b_q$. In the usual case, where these elements are to be reduced to equal standard deviations before combination, it may be shown that

$$r_{(a_1+a_2+\ldots+a_p)\,(b_1+b_2+\ldots+b_q)} = \frac{\sqrt{pq}\,\bar{r}_{ab}}{\sqrt{1+(p-1)\bar{r}_{aa}}\,\sqrt{1+(q-1)\bar{r}_{bb}}} \quad \ldots\ldots(1)^1,$$

where the term on the left is the required correlation between the sum of the a's and the sum of the b's; $\bar{r}_{aa}$ is the mean correlation between all different pairs of a's; $\bar{r}_{bb}$ is that between all different pairs of b's; and $\bar{r}_{ab}$ is that between all different pairs of an a and a b.

This equation may sometimes more conveniently be written in the form

$$r_{(a_1+a_2+\ldots+a_p)\,(b_1+b_2+\ldots+b_q)} = \frac{S(r_{ab})}{\sqrt{p+2S(r_{aa})}\,\sqrt{q+2S(r_{bb})}} \quad \ldots\ldots(2)^1,$$

where $S(r_{aa})$ is the sum of all correlations between different pairs of a's; $S(r_{bb})$ is that between all different pairs of b's; and $S(r_{ab})$ is that between all different pairs of an a and a b.

From (2) we can readily get the case where the standard deviations of the a's or of the b's are unequal. We have only to multiply every r contained in equation (2) by the standard deviations of the two series entering into this r. The whole equation then becomes

$$r_{(a_1+\ldots+a_p)\,(b_1+\ldots+b_q)} = \frac{S(\sigma_a\sigma_b r_{ab})}{\sqrt{S(\sigma_a{}^2)+2S(\sigma_a\sigma_a r_{aa})}\,\sqrt{S(\sigma_b{}^2)+2S(\sigma_b\sigma_b r_{bb})}} \quad \ldots\ldots(3)^1,$$

where $S(\sigma_a{}^2)$ is the sum of the squared standard deviations of the a's; $S(\sigma_b{}^2)$ is that of the b's; $S(\sigma_a\sigma_a r_{aa})$ is the sum of the correlations between all different pairs of a's, each such correlation being multiplied by the σ's of the two a's concerned; and $S(\sigma_b\sigma_b r_{bb})$ and $S(\sigma_a\sigma_b r_{ab})$ are analogous values.

Sometimes it may be desired to multiply some or all of the variables $a_1, a_2, \ldots a_p, b_1, b_2, \ldots b_q$ by any constants $n_1, n_2, \ldots n_p, m_1, m_2, \ldots m_q$. To get this, we have only to multiply every σ in equation (3) by the n (or m) belonging to the same a (or b). Whether these constants be positive or negative integers or fractions, we obtain

$$r_{(n_1a_1+\ldots+n_pa_p)\,(m_1b_1+\ldots+m_qb_q)} = \frac{S(nm\sigma_a\sigma_b r_{ab})}{\sqrt{S(n^2\sigma_a{}^2)+2S(nn\sigma_a\sigma_a r_{aa})}\,\sqrt{S(m^2\sigma_b{}^2)+2S(mm\sigma_b\sigma_b r_{bb})}} \quad \ldots(4)^1,$$

where the meaning of the terms is evident from the above.

[1] Appendix, § 1. An illustrative example is worked out in § 2 of the Appendix.

There is no difficulty now in determining the proportions in which the elements have to be combined, in order to make the correlation of the sum (or difference) as large or as small as possible. Take the simple case of any three variables a_1, a_2 and b. Suppose that it is required to determine n so as to make $r_{(na_1+a_2)(b)}$ a maximum or a minimum, n being any positive or negative integer or fraction. This is done by evaluating the r in accordance with the above (4), and then differentiating[1]. We find that the equation

$$n = \frac{\sigma_{a_2}}{\sigma_{a_1}} \cdot \frac{r_{a_1 b} - r_{a_2 b} \cdot r_{a_1 a_2}}{r_{a_2 b} - r_{a_1 b} \cdot r_{a_1 a_2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)^2$$

makes $r_{(na_1+a_2)(b)}$ a maximum when $r_{a_2 b}$ is greater than $r_{a_1 b} \cdot r_{a_1 a_2}$, and a minimum when $r_{a_2 b}$ is less than $r_{a_1 b} \cdot r_{a_1 a_2}$. In the former case the minimum, and in the latter case the maximum, is given by

$$n = \pm \infty \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6),$$

where the sign to precede ∞ is the opposite to that of $r_{a_1 b}$ for the minimum, and the same as that of $r_{a_1 b}$ for the maximum.

By putting any of these values of n into any of the above equations (1) to (4), we get the corresponding values of $r_{(na_1+a_2)(b)}$. The latter becomes simply $r_{a_1 b}$ when $n = +\infty$, and $-r_{a_1 b}$ when $n = -\infty$. When n takes the value shown in (5), then

$$r_{(na_1+a_2)(b)} = \pm \sqrt{\frac{r^2_{a_1 b} r^2_{a_2 b} - 2 r_{a_1 b} \cdot r_{a_2 b} \cdot r_{a_1 a_2}}{1 - r^2_{a_1 a_2}}} \quad \ldots\ldots\ldots (7)^3,$$

taking the plus sign when $r_{a_2 b}$ is greater than $r_{a_1 b} \cdot r_{a_1 a_2}$, and vice versa.

Turn, lastly, to the case where the number of series to be added

[1] The required 'relative' maximum is obtained by observing the variations of the sign of the first differential (see Serret, *Cour de Calcul différential*, § 146). In many cases it coincides with an 'absolute' maximum, and is therefore among the solutions got from putting the differential = zero.

[2] Here I find myself, as often before, partly anticipated by the correlational researches of Udny Yule. He showed some time ago (as was kindly pointed out to me by Mr Webb) that equation (5) gives the maximum value of $r_{(na_1+a_2)(b)}$ if we disregard sign, which we usually can do (*J. Roy. Statist. Soc.*, 1906, 199). Moreover, even the more general case of the maximum value of $r_{(n_1 a_1 + \ldots + n_p a_p)(b)}$ appears readily soluble by the theory of partial correlations of the same investigator (see his *Introduction to the Theory of Statistics*, 1912, chap. XII. section 16). By the aid of the present theorems, the still more general case of the maximum of $r_{(n_1 a_1 + \ldots + n_p a_p)(m_1 b_1 + \ldots + m_q b_q)}$ can be obtained, though sometimes through rather complicated differentiation.

[3] This root expression is termed by Yule 'R' and has been shown by him to possess considerable statistical importance (see his *Introduction to the Theory of Statistics*, 243–4).

together is infinite. In actual experience, of course, only a comparatively small number can possibly be given; we can only observe the correlation of a with the sum (or average) of q b's where q is finite. But we may well wish to know the correlation of a with the sum or average of an infinite number of b's that are *similar*, in the sense of having the same average correlation with a and also with one another. From (1) we easily get

$$r_{(a)\,(b_1+b_2+\ldots\text{ to infinity})} = r_{(a)\,(b_1+\ldots+b_q)} \frac{\sqrt{1+(q-1)\,\bar{r}_{bb}}}{\sqrt{q\,.\,\bar{r}_{bb}}} \quad \ldots\ldots(8)^1,$$

and in the same way

$$r_{(a_1+a_2\ldots\text{ to inf.})\,(b_1+b_2\ldots\text{ to inf.})} = r_{(a_1+a_2+\ldots+a_p)\,(b_1+b_2+\ldots+b_q)} \frac{\sqrt{1+(p-1)\,\bar{r}_{aa}}\sqrt{1+(q-1)\,\bar{r}_{bb}}}{\sqrt{p\,.\,\bar{r}_{aa}}\sqrt{q\,.\,\bar{r}_{bb}}} \quad \ldots(8a)^1,$$

where $\bar{r}_{aa}$ denotes, as before, the average correlation of the p a's with one another, and $\bar{r}_{bb}$ is analogous. We can give to p or to q any values we please, but in practice they are most frequently taken as $= 2$.

If we replace $r_{(a_1+\ldots+a_p)\,(b_1+\ldots+b_q)}$ in $(8a)$ by its value as given in (1), we get the simple result

$$r_{(a_1+\ldots\text{ to inf.})\,(b_1+\ldots\text{ to inf.})} = \frac{\sqrt{\bar{r}_{ab}}}{\sqrt{\bar{r}_{aa}\,.\,\bar{r}_{bb}}} \quad \ldots\ldots\ldots\ldots(9)^1,$$

where $\bar{r}_{ab}$ means, as before, the average correlation between the p a's and the q b's, and the remaining terms have the same significance as in $(8a)$.

The above three formulae are *exact* for any values whatever of the a's and the b's, so long as the above 'similarity' holds good, that is, the average correlation between the a's, between the b's, and between the a's and b's is the same on the left side of the equation as on the right. The whole real difficulty lies in the question as to how far our observed series can be expected to give 'similar' correlations to those of the infinite series that we may have in mind.

To begin with, this expectation of similarity is obviously precarious in proportion to the size of the 'probable error' of $\frac{\sqrt{\bar{r}_{ab}}}{\sqrt{\bar{r}_{aa}\bar{r}_{bb}}}$. If $\bar{r}_{aa}$ and $\bar{r}_{bb}$ are very small, this 'probable error' is usually very large, so that the determination of $r_{(a_1+\ldots\text{ to inf.})\,(b_1+\ldots\text{ to inf.})}$ becomes illusory. In practice, I have found it very desirable that $\bar{r}_{aa}$ and $\bar{r}_{bb}$ should amount to ·70 to ·80, and indispensable that they should be at least ·40 to ·50.

[1] Appendix, § 4.

Further, only in certain cases does the basal conception, that of an infinite number of series giving 'similar' correlations to the series actually observed, possess a useful scientific significance. One such case is where $a_1, a_2, \ldots$ are several measurements of one and the same variable, and differ from one another merely owing to random errors; similarly, as regards $b_1, b_2, \ldots$. The formula for this case has been given in former papers, and will be found to be perfectly corroborated by the present equations (8 *a*) and (9)[1].

Another important case is where the mean of an infinite number of values denotes, not any individual value, but some *collective characteristic* (as average, dispersion, etc.) of the ideal frequency distribution of infinitely numerous results, for which certain conditions have remained constant while other influences have varied in a random manner. Under this heading would come all tests of mental powers, as discussed above.

Or again, the mean of infinitely numerous values is often of use even when it represents some magnitude plus a definite error. The measurements of anything will be liable to disturbances which are in part random but in part constantly biased in some particular direction. On taking the mean of an infinite number of such measurements, the random errors will be completely eliminated; the constant error will persist, indeed, but now it will be cleared of complication and amenable to special treatment[2]. This applies equally to collective characteristics; for these, too, are liable to distortion by constant influences—whether downright errors, or merely tendencies irrelevant to the point at issue—which must be treated just like the constant errors that occur in measuring an individual.

The chief difficulty arises when, in addition or not to any random and any constant errors, there are slow systematic changes, like those produced in mental performances by exercise and fatigue. Here, the application and interpretation of (8*a*) and (9) will need care[3]. But

[1] This *Journal*, 1910, III. p. 275; *Amer. Journ. of Psychol.* 1904, XV. p. 290. Some critics have wrongly said that the original formula was subsequently changed. All that has since been done is to render it much more general and to improve its application.

[2] See this *Journal*, 1910, III. 279.

[3] Under such circumstances, (8*a*) and (9) will sometimes yield values which are impossible, not lying between +1 and −1. This does not show that the equation is wrong, but only that the case of which they treat—an infinite number of *a*'s and *b*'s presenting on an average the same sized inter-correlations as the *a*'s and *b*'s observed—cannot occur with the given values (usually, owing merely to their errors of sampling). This impossibility can be traced to the logical requirement that, whatever series 1, 2, and 3 may denote, r_{23} must necessarily lie between the limits: $r_{12}r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 - r_{12}^2 r_{13}^2}$ (see Yule, *ibid.* p. 246). The converse fact is worth remembering: whenever (8*a*) and (9)

nearly always, the values obtained will be the sums or averages of several single observations; for instance, a performance of memory will not be measured by the correctness of one or even two answers only, but by that of a considerable number. And these single observations can easily be arranged in two groups, a_1 and a_2, whose difference with respect to the slow changes is negligible[1]. Hereupon, (8a) and (9) become applicable with the same signification as in the two preceding paragraphs.

Further uses of all the equations given above will readily suggest themselves to any one familiar with correlational work. For example, (1) furnishes at once "the increase of correlation between two different characters, to be obtained by increasing the number of measurements," for which the writer has given a formula in a previous paper[2]; in fact, all the main formulae more or less elaborately demonstrated in that paper turn out to be immediate corollaries of our simply obtained (1)[3].

Many results of other writers can be got with equal facility. For instance, a few months ago Woodworth proved that the correlation of $(x+y)$ with x is $1/\sqrt{2}$[4]. But if in (1) we put $a_1 = b = x$ and $a_2 = y$, we have immediately

$$r_{(x+y)(x)} = \frac{\sqrt{2} \times \frac{1}{2}}{\sqrt{1+0}\sqrt{1}} = 1/\sqrt{2}.$$

Further instances have appeared above in the values that had been previously reached in various manners by Udny Yule.

To sum up, the cases where we need the correlations of sums, averages, and differences are numerous and important. The above simple formulae both aid in obtaining them, and also show that the customary replacement of the correlation of averages by the average of correlations [namely, $r_{(a_1+\ldots+a_p)(b_1+\ldots+b_q)}$ by $\bar{r}_{ab}$] cannot under any circumstances lead to the right result. It should also be noticed that all the above formulae are in themselves not approximate but exact; any inaccuracy in the conclusions got through their use can only mean

give values between $+1$ and -1, then it is possible for an infinite number of a's and b's to give on an average the same sized inter-correlations as those already observed.

[1] Usually it will be sufficient to pool the odd single values for a_1, and the even ones for a_2.

[2] This *Journal*, 1910, III. 281.

[3] Hence it was a mistake in that paper to suppose them at all independent of one another (see top of p. 282).

[4] *Psychol. Rev.* 1912, XIX. 113. He assumed, as we do here, that the standard deviations of x and y are equal.

that the terms in the formulae are untruly represented by the empirical data employed or by the theoretical interpretations given. Also, the formulae do not involve either 'normality' in the distribution of the variables or 'linearity' in the correlations[1]; they make no assumptions beyond those of logic.

APPENDIX.

§ 1. Let the two series of variables to be summed be denoted by $a_1, a_2, \ldots a_p$ and $b_1, b_2, \ldots b_q$, each being measured from its own mean and consisting of N cases. Let these variables be multiplied respectively by the constants $n_1, n_2, \ldots n_p$ and $m_1, m_2, \ldots m_q$. Let the required correlational coefficient between

$$n_1a_1+n_2a_2+\ldots+n_pa_p \text{ and } m_1b_1+m_2b_2+\ldots+m_qb_q$$

be denoted by $r_{(n_1a_1+\ldots+n_pa_p)(m_1b_1+\ldots+m_qb_q)}$.

Then the numerator of this coefficient will evidently

$$=\sum_1^N\left[\sum_1^p S(na)-\frac{1}{N}\sum_1^N\sum_1^p S(na)\right].\sum_1^N\left[\sum_1^q S(mb)-\frac{1}{N}\sum_1^N\sum_1^q S(mb)\right],$$

where S denotes summation of the p or q different variables, and Σ denotes summation of the N different cases.

This expression reduces to

$$NS_{st}(n_sm_t\sigma_{a_s}\sigma_{b_t}r_{a_sb_t}),$$

where $r_{a_sb_t}$ is the correlation between a_s and b_t, the two σ's are the standard deviations of these two variables, and s and t take all values from 1 up to p and q respectively. Putting similar terms in the denominator, we get the required coefficient,

$$r_{(n_1a_1+\ldots+n_pa_p)(m_1b_1+\ldots+m_qb_q)}=\frac{S_{st}(n_sm_t\sigma_{a_s}\sigma_{b_t}r_{a_sb_t})}{\sqrt{S^*_{st}(n_sn_t\sigma_{a_s}\sigma_{a_t}r_{a_sa_t})}\sqrt{S^*_{st}(n_sm_t\sigma_{b_s}\sigma_{b_t}r_{b_sb_t})}} \quad\ldots(10),$$

the stars indicating that these sums include the cases where a_s and b_s denote the same variables as a_t and b_t respectively,

$$=\frac{S_{st}(n_sm_t\sigma_{a_s}\sigma_{b_t}r_{a_sb_t})}{\sqrt{S_s(n_s^2\sigma_{a_s}^2)+2S_{st}(n_sn_t\sigma_{a_s}\sigma_{a_t}r_{a_sa_t})}\sqrt{S_s(m_s^2\sigma_{b_s}^2)+2S_{st}(m_sm_t\sigma_{b_s}\sigma_{b_t}r_{b_sb_t})}}\quad\ldots(11),$$

where a_s and b_s always denote different values from a_t and b_t respectively.

When each n and m is equal to unity, and the σ's of the a's are equal to one another, and similarly the σ's of the b's, then on reduction

$$r_{(a_1+\ldots+a_p)(b_1+\ldots+b_q)}=\frac{S_{st}(r_{a_sb_t})}{\sqrt{p+2S_{st}(r_{a_sa_t})}\sqrt{q+2S_{st}(r_{b_sb_t})}} \quad\ldots\ldots\ldots\ldots(12),$$

[1] Although non-linearity cannot affect the relations given in this paper between the r's, it can, of course, influence the significance of these r's. But to say that extreme non-linearity makes r meaningless overlooks the distinction between correlation and dependence (Yule, *ibid.* p. 174) and the extraordinarily wide-reaching statistical properties of r in any case.

which may, sometimes more conveniently, be written as

$$r_{(a_1+\ldots+a_p)(b_1+\ldots+b_q)}=\frac{\sqrt{pq}\,.\,\bar{r}_{ab}}{\sqrt{1+(p-1)\,\bar{r}_{aa}}\sqrt{1+(q-1)\,\bar{r}_{bb}}} \quad \ldots\ldots\ldots\ldots(13),$$

where $\bar{r}_{ab}$ denotes the mean correlation between the a's and b's, and $\bar{r}_{aa}$ and $\bar{r}_{bb}$ are analogous values.

§ 2. An illustrative example of the above.

Suppose that we want the standard deviation of a_2 to be double that of a_1, and that of b_2 to be treble that of b_1, while, further, b_2 is to be multiplied by $-\cdot3$. Then, by (11), the required coefficient

$$r_{(a_1+a_2)(b_1-\cdot3b_2)}=\frac{\sigma_{a_1}\sigma_{b_1}r_{a_1b_1}-\cdot3\sigma_{a_1}\sigma_{b_2}r_{a_1b_2}+\sigma_{a_1}\sigma_{b_1}r_{a_2b_1}-\cdot3\sigma_{a_2}\sigma_{b_2}r_{a_2b_2}}{\sqrt{\sigma_{a_1}^2+\sigma_{a_2}^2+2\sigma_{a_1}\sigma_{a_2}r_{a_1a_2}}\sqrt{\sigma_{b_1}^2+\cdot09\sigma_{b_2}^2-\cdot6\sigma_{b_1}\sigma_{b_2}r_{b_1b_2}}},$$

and, dividing out by $\sigma_{a_1}\sigma_{b_1}$,

$$=\frac{r_{a_1b_1}-\cdot3r_{a_1b_2}+r_{a_2b_1}-\cdot3r_{a_2b_2}}{\sqrt{5+4r_{a_1a_2}}\sqrt{1\cdot81-1\cdot8r_{b_1b_2}}}.$$

§ 3. To find the maximum and minimum of $r_{(na_1+a_2)(b)}$ for varying n.

By (11) the above coefficient

$$=\frac{n\sigma_{a_1}\sigma_b r_{a_1b}+\sigma_{a_2}\sigma_b r_{a_2b}}{\sqrt{n^2\sigma_{a_1}^2+\sigma_{a_2}^2+2n\sigma_{a_1}\sigma_{a_2}r_{a_1a_2}}\sqrt{\sigma_b^2}},$$

so that its first differential, on some reduction,

$$=\frac{n\sigma_{a_1}^2\sigma_{a_2}(r_{a_1a_2}r_{a_1b}-r_{a_2b})+\sigma_{a_1}\sigma_{a_2}^2(r_{a_1b}-r_{a_1a_2}r_{a_2b})}{(n^2\sigma_{a_1}^2+\sigma_{a_2}^2+2n\sigma_{a_1}\sigma_{a_2}r_{a_1a_2})^{\frac{3}{2}}} \quad \ldots\ldots\ldots\ldots(14).$$

Now, take first the case that $r_{a_2b}>r_{a_1a_2}r_{a_1b}$.

Let n vary continuously from $-\infty$ to $+\infty$. At the start, the above differential $=\frac{r_{a_2b}-r_{a_1b}r_{a_1a_2}}{\infty^2}$, which is positive; its denominator being so always, the differential remains positive until its numerator passes through zero, whereupon it becomes and remains negative. Hence, the required maximum occurs when its numerator $=0$, namely, when

$$n=\frac{\sigma_{a_2}}{\sigma_{a_1}}\frac{r_{a_1b}-r_{a_2b}r_{a_1a_2}}{r_{a_2b}-r_{a_1b}r_{a_1a_2}} \quad \ldots\ldots\ldots\ldots(15).$$

When $n=-\infty$ and $=+\infty$, $r_{(na_1+a_2)(b)}$ reduces to $-r_{a_1b}$ and to $+r_{a_1b}$ respectively; clearly, whichever of these two values is negative is the required minimum.

Take, next, the case that $r_{a_2b}<r_{a_1b}r_{a_1a_2}$.

By similar reasoning to the above, $r_{(na_1+a_2)(b)}$ can be shown to attain its maximum at the positive one of the two values, r_{a_1b} and $-r_{a_1b}$; also to attain its minimum when n becomes the value given in (15).

§ 4. To find the value of $r_{(a_1+a_2+\ldots \text{ to infinity})(b_1+b_2+\ldots \text{ to infinity})}$.

The required coefficient, using (13),

$$=\frac{\sqrt{\infty \times \infty}\,\bar{\bar{r}}_{ab}}{\sqrt{1+(\infty-1)\bar{\bar{r}}_{aa}}\sqrt{1+(\infty-1)\bar{\bar{r}}_{bb}}}=\frac{\bar{\bar{r}}_{ab}}{\sqrt{\bar{\bar{r}}_{aa}\,.\,\bar{\bar{r}}_{bb}}},$$

where $\bar{\bar{r}}_{ab}$ indicates the mean correlation of the infinitely numerous a's with the infinitely numerous b's, and $\bar{\bar{r}}_{aa}$ and $\bar{\bar{r}}_{bb}$ have analogous meanings.

This becomes, on multiplying it by the left-hand term and dividing by the right-hand term of (13),

$$r_{(a_1+a_2+\ldots+a_p)(b_1+b_2+\ldots+b_q)}\cdot\frac{\sqrt{1+(p-1)\bar{r}_{aa}}\sqrt{1+(q-1)\bar{r}_{bb}}}{\sqrt{pq\,.\,\bar{r}_{ab}}}\cdot\frac{\bar{\bar{r}}_{ab}}{\sqrt{\bar{\bar{r}}_{aa}\,.\,\bar{\bar{r}}_{bb}}},$$

where p and q have any values we please; and, if $\bar{r}_{aa}=\bar{\bar{r}}_{aa}$, $\bar{r}_{bb}=\bar{\bar{r}}_{bb}$, $\bar{r}_{ab}=\bar{\bar{r}}_{ab}$, the required coefficient

$$=r_{(a_1+a_2+\ldots+a_p)(b_1+b_2+\ldots+b_q)}\frac{\sqrt{1+(p-1)\bar{r}_{aa}}}{\sqrt{p\bar{r}_{aa}}}\,.\,\frac{\sqrt{1+(q-1)\bar{r}_{bb}}}{\sqrt{q\bar{r}_{bb}}}\quad\ldots\ldots\ldots(16).$$

(*Manuscript received* 5 *January* 1913.)