CORRELATION CALCULATED FROM FAULTY DATA.

By C. SPEARMAN.

I. *A correlational formula eliminating "accidental" errors of measurement.*

A FEW years ago I called attention to the fact that the apparent degree of correspondence between any two series of measurements is largely affected by the size of the "accidental" errors in the process of measurement[1]. It was pointed out that this disturbance is not in the least bettered by making the series longer. As a remedy, a correction formula was proposed, based on the idea that the size of these "accidental" errors can be measured by the size of the discrepancies between successive measurements of the same things[2].

Now, all experimenters seem to be unanimous in finding that such discrepancies are liable to be startlingly large. The importance of the point is therefore established. For an estimate of the correlation between two things is generally of little scientific value, if it does not depend unequivocally on the nature of the things, but just as much on the mere efficiency with which they happen to have been measured.

But nevertheless the formula has met with much opposition. When first published, some eminent authorities at once declared it

[1] This illusion is, of course, just as bad when the correspondence is judged by general impression instead of by coefficients; in fact, worse, as then no correction is possible.

[2] *Am. J. Psych.* Vol. xv. 1904.

to be mathematically incorrect. This attitude appears now to have been abandoned in favour of a more moderate line of resistance. The formula is allowed to be true for really "accidental" errors; but it is urged that, in psychology at all events, the discrepancies between successive measurements often cannot properly be termed "accidental," but may arise from the fact that the second later measurement does not deal with the same function as the earlier one, owing to the modifications introduced by practice, fatigue, etc. Hence, the correction formula becomes invalid; indeed, as it throws together two different functions, it is even meaningless.

Some idea of this sort will be found already in my original paper[1]; and it was quickly and clearly emphasized by Wissler[2]. But the crucial point was reached when Udny Yule gave a new and much simpler proof of the formula, putting its validity beyond further cavil, but showing in the plainest light the assumptions on which it is based[3]. These are, that the errors of measurement are not correlated with one another or with either of the series measured. Clearly, such assumptions are far from carrying conviction *a priori*. And, finally, Dr W. Brown has furnished some actual experimental instances of their invalidity, as well as some interesting theoretical discussion[4].

One remedy that has been suggested is to make the measurements so efficient, that the correction will not be needed. But how are we to tell whether our measurements really are efficient enough, except by trying with the correction formula? The suggestion is like telling a man to brush his coat until it is clean but never look whether it is so. Also, the half measure has been advised, of making the successive measurements prescribed by the formula, noting whether there is much

[1] *Ibidem*, pp. 254, 255.

[2] *Science*, Vol. XXII. pp. 309 ff.

[3] In a private letter sent to me in October, 1908; his proof is attached in appendix *e*. My own far less elegant proof makes, I think, a little less extensive assumptions; but the difference is unimportant. He has also had the great kindness to look through this paper and to suggest numerous criticisms, to which it is very much indebted. I have, further, the pleasure of acknowledging helpful criticisms and remarks from Dr Nunn, Mr Sheppard, Mr Burt, and Dr Betz.

[4] The experimental evidence brought by him at the Geneva Congress cannot be admitted. Let x denote the true measurement, x_1 and x_2 the first and second measurements actually obtained, d_1 and d_2 their errors; for x he substituted x_1, which $=x+d_1$, thus illicitly bringing in d_1, the very quantity in question. He, further, advanced the view, that the formula ceases to hold good whenever ability (measured by x) is correlated with "variability" (measured by $x_1 \sim x_2$); but he appears to have based this on the mistaken notion, that correlation with $x_1 \sim x_2 = d_1 \sim d_2$ proves correlation with either d_1 or d_2. In a later paper, however, he brings more satisfactory experimental evidence; also, he seems to give up the view about variability (*Biometrika*, Vol. VII. p. 352).

discrepancy, but not proceeding to use the formula for evaluating the effect of the discrepancy on the coefficient. This is equally futile; for the seriousness of the discrepancy can only be gauged by its effect on the coefficient.

The difficulty must be met more drastically. To begin with, we may note that in a large number of cases, the questioned assumptions are legitimate enough, for instance, in the measurement of physical objects. Unfortunately, this is not very helpful; for in most physical measurements the errors are exceedingly small; and the correction formula has proved that such very small errors affect the coefficient too little to demand, for most purposes, elimination. This elimination is needed rather in such sciences as psychology and sociology; in testing, for instance, a person's power of bisecting a line, we find that almost every successive trial yields a distinctly different result. But it is just in such sciences that the assumptions become most dangerous.

Here, however, we can fall back on the universal and invaluable device of analysing, in thought, such variations into components of two kinds. Firstly, there are the variations of a *regular*, generally a continuously progressive character. These demand and admit of investigation, explanation, and, in large degree, control. In our above example, we should find that the accuracy of bisection was being increased by practice, diminished by fatigue, etc. It is with regard to this kind of component that Wissler, Yule and Brown are unquestionably right in calling attention to the dubious validity of the old formula's assumptions; it is certainly more than hazardous to assume, for instance, that fatiguability in one performance is uncorrelated with ability in that performance, or with fatiguability in another performance. At the same time, recent research seems to indicate that such correlations are far smaller than popularly supposed; I am not myself aware of any conclusions arrived at by means of the old formula which would probably be upset on taking such correlations into account.

Secondly, superposed on the above regular variations, we find others of such a discontinuously shifting sort, that investigation, explanation, and control are almost baffled. Hence, we call these by some such name as "accident." Of course, exceptional cases may be conceived where the line demarcating the accidental from the regular components becomes obscure, but in the immense majority of cases it is perfectly clear and usable.

Now, it is the superposed accident that the present paper attempts to eliminate, herein following the custom of all sciences, one that appears to be an indispensable preliminary to getting at nature's laws. This elimination of the accidents is quite analogous to, and serves just the same purpose as, the ordinary processes of "taking means" or "smoothing curves." The underlying regular variations, on the other hand, do not in general require elimination, but only determination. Every mental performance, for instance, must necessarily be at some stage or other of practice or fatigue; every stage is equally "true," and forms an equally legitimate subject of investigation. All that can reasonably be demanded of a formula is to produce the coefficient for some definite stage, and we will here choose the average stage during the period of measurement.

But, as regards this second or "accidental" component to be eliminated, the assumptions as made by the old formula seem to possess an exceedingly wide validity. This paper proposes, then, to suggest a new formula, or rather to raise the old one to a higher generality, such as to involve these assumptions only as regards the "accidental" components, where they are legitimate; not as regards the regular components, where they may be called in question. We will take, however, the precaution of discussing the circumstances under which the assumptions may conceivably become invalid even as regards the "accidents."

The method is as follows. Let each individual be measured several times with regard to any characteristic to be compared with another. And let his measurements be divided into several—usually two—groups. Then take the average of each group; this we will term the "group average." *The division into groups is to be made in such a way, that any differences between the different group averages (for the same individual) may be regarded as quite "accidental."* It is further desirable, that the sum total of the accidental variations of all the individuals should be not very unequal in the different groups; ordinarily, this will occur without further trouble, but in any case it can be arranged[1].

Such a division seems feasible in most psychological and sociological work. A test of verbal memory, for instance, might well consist of memorizing twenty series of words (exclusive of some preliminary series

[1] It must be noted that we can rarely assume the "accidents" in the measurements of a single individual eventually to cancel one another on taking an average. This would postulate a much larger number of measurements than usually attainable. And as the errors thus introduced for each individual are squared in calculating the coefficient, they would not tend mutually to cancel one another when added together, but would exercise a definite and often large bias.

for "warming up"). Then series 1, 3, 5, ... 19 would suitably furnish one group, while the even numbers gave the other. Any discrepancy between the averages of the two groups might, as a rule, be regarded as practically all due to the "accidents."

Quite a small number of measurements will suffice when they extend over a brief period or when the variation may be assumed to proceed uniformly. If the variation is sensibly uniform—an assumption always valid if the period is sufficiently brief,—three measurements will be enough; one group can consist of 1 and 3; the other can be represented by 2. If four groups are available, one can consist of 1 and 4, the other of 2 and 3. When there is no lapse of time between the measurements, each one of them may replace a group; a common instance is that of measurements of children which consist in classifications or orders of merit derived from the general impressions of their teachers.

The result of this division into group averages (or classifications, etc.) is that each of these has two components, the average underlying "regular" measurement, and the average superposed "accidental" disturbance. And as regards the latter component, there appears no reasonable objection to assuming it to be uncorrelated both with the accidental components of the other group averages and with the underlying "regular" measurements. On making this assumption, we obtain the following equation (for proof, see app. c):

$$\mathbf{r}_{xy} = \mathbf{r}_{x[p],y[q]} \cdot \sqrt{\frac{1+(\mathbf{p}-1)\,\mathbf{r}_{x[1],x[1]}}{\mathbf{p}\cdot\mathbf{r}_{x[1],x[1]}} \cdot \frac{1+(\mathbf{q}-1)\,\mathbf{r}_{y[1],y[1]}}{\mathbf{q}\cdot\mathbf{r}_{y[1],y[1]}}} \quad \ldots (\mathrm{I}),$$

where $r_{x[1],x[1]}$ denotes the average correlation between the single groups averages for x; $r_{y[1],y[1]}$ does the same for y; $r_{x[p],y[q]}$ denotes the correlation of the average of p group averages for x with the average of q for y; and r_{xy} is the desired correct correlation, i.e., that between the average values of the underlying regular measurements of x and y respectively.

Here, p and q may have any chosen values, but it is best to make them = the total number of groups formed of x and y respectively; for thus the greatest possible approximation to the correct value of the correlation is obtained directly, and the least possible influence is left to the factor expressed in (I) as a square root. The number of groups should generally be *two* only, since thus concentrating the measurements into few groups facilitates the complete elimination of all the "non-accidental" discrepancies between the group averages. The formula then becomes:

$$\mathbf{r}_{xy} = \mathbf{r}_{x[2], y[2]} \cdot \sqrt{\frac{1 + \mathbf{r}_{x[1], x[1]}}{2\mathbf{r}_{x[1], x[1]}} \cdot \frac{1 + \mathbf{r}_{y[1], y[1]}}{2\mathbf{r}_{y[1], y[1]}}}.$$

It may be noted that, by putting $p = q = 1$, we return to my original formula[1], the only difference being the present improved method of constituting the groups[2].

If $r_{x[1], x[1]}$ or $r_{y[1], y[1]}$ is unknown, there is no resource but putting it equal to 1, as is tacitly done in the Bravais formula as ordinarily calculated. The result, however, as in that formula, is not the correct coefficient, but merely the minimum which the correct coefficient cannot fall short of.

The "probable error" of sampling is, approximately,

$$= \cdot 6745 \frac{1 - \mathbf{r}^r_{x[p], y[q]}}{\sqrt{n}} \cdot K \quad \text{....................(II)},$$

here p and q denote the total number of groups of x and y respectively; K denotes the square root in (1), and n is the number of individuals. Thus, the p.e. of the correct coefficient = the p.e. of the Bravais coefficient calculated in the ordinary manner divided by the ratio of the latter to the correct coefficient[3].

Let us now consider the possible exceptions even to the above curtailed assumptions. For simplicity, we will take the case of x and y each furnishing two group averages, which we will term x_a, x_b, y_a and y_b. The assumptions made are, then, that the "accidental" components in these four terms respectively are uncorrelated with one another and with x and y.

Take, first, the possibility of correlation of the accidents of x_a with those of x_b. This could only mean that the accidents had a general bias in favour of some individuals as compared with others. Then, clearly, the formula will give the correlation, not between the true values of x and y, but, in general, between these values as biassed. And it could hardly be expected to do more. Such a bias can only be eliminated by improving the methods of obtaining the data.

Take next any correlation of the accidents in x_a with x or y. The above again holds good, except that in this special case statistics do

[1] *Am. J. Psych.* Vol. XVIII. p. 168.

[2] It is often useful to choose two *different* values for p and also for q, and to see whether they lead to two very discrepant values for r_{xy}. A large discrepancy indicates something wrong in the assumptions or elsewhere.

[3] It thus coincides with the approximation suggested by me to Mr Burt and published by him in the *British Journal of Psychology*, Vol. III. p. 111, 1909.

furnish a possible means of remedying the faultiness of the data, namely, Yule's formula for eliminating irrelevant factors. This will be discussed in the next section.

It remains to consider the possibility of correlation of the accidents in some of the x measurements with those in some of the y measurements. In experimental psychology, for instance, it is not uncommon for each individual to be tested separately, and for each test in x to be accompanied by a test in y. Suppose, now, any individual to be accidentally indisposed: his results for both x and y will be accidentally depressed; the same will occur, more or less, for the other individuals; hence arises a spurious correlation between x and y. It may, however, easily be avoided; let the accidents in the pth tests of x and y be called d_p and e_p; we need only arrange so as to omit $d_p . e_p$ from $s(d_{ab} . e_{ab})$, see appendix *a*. Of course, it is advisable, where possible, to get better data to start with; in the above case, it might be practicable to test x and y on separate occasions; or means might be devised of ascertaining when the individuals are indisposed, etc.

A point to be noticed about this formula is that, like the former one proposed by me, it will occasionally produce coefficients greater than unity. Some authors have strongly objected to this[1]. But the objection would only be justifiable if the coefficient pretended to be perfectly accurate. At most, it is only the true coefficient plus the error due to testing a limited sample instead of the whole class; the general magnitude of such an error is indicated by the so-called "probable error." And though a true coefficient cannot exceed unity, there is no reason why a coefficient plus an error should not do so. In such case, of course, the coefficient must be taken as $= 1$, this being its most probable value.

In view of the easy statistical elimination of the accidental errors, it might be thought no longer necessary to make long careful measurements. But this would be a grave mistake; for as seen from equation (II), such accidents swell the correct coefficients probable error. Hence the function of the formula here proposed is by no means to replace accurateness of original data, but on the contrary to emphasize the necessity of such accurateness, to estimate the degree of its realization, and only in the last instance to supplement its defectiveness.

It should be noted that the above proposed elimination of accidental variations bases itself on the original Bravais coefficient, this

[1] Cf. Pearson, *Biometrika*, Vol. III. p. 160. His other principal criticisms will be discussed in Section VI.

appearing to me the most generally satisfactory one hitherto invented. But at the same time, it must be admitted that this coefficient has many weaknesses and that other coefficients have many advantages[1]. It must, however, be demanded of these other coefficients, no less than of the Bravais one, that they be analogously modified so as to eliminate the effect of the accidents.

II. *Yule's Correction for Irrelevant Factors.*

Difficulties are not yet over. So far, we have only eliminated from our coefficient the effect of "accidents" in our measurements. But our data are no less liable to be affected by non-accidental or general tendencies. To avoid such disturbing tendencies is primarily, of course, the concern of the method of obtaining the data. There is, however, one important case where faultiness of the data in this respect admits of subsequent correction statistically; it is the case where the undue tendency consists in the character under estimation being influenced by some *irrelevant factor*. These arise from the fact that the scientific conquest of nature is essentially achieved by artificially simplifying her processes; the factor that we happen to be investigating is allowed to vary, while the remaining factors are kept as constant as possible, their effect being regarded as irrelevant for the purpose in hand. Suppose, for instance, we wanted to ascertain the correlation between ability for composition and for mathematics. It would not be legitimate to pool together schools devoting different amounts of time to instruction. Otherwise, the fact that the children high in the one subject tended to be high in the other also might merely mean that these children came from schools giving more instruction in both subjects than the other schools. Such difference of amount of instruction constitutes a factor irrelevant and disturbing to our purpose; it requires elimination.

Unfortunately, these irrelevant factors are innumerable and ubiquitous. The shallow, or careless worker, or one trying to make statistical calculations replace special knowledge of the subject, is at their mercy. Even the most thorough and competent investigator, after completing his experiments, working out his correlations, and eliminating the influence of accidents, will still have many a misgiving about the irrelevancies; some factor may now occur to him of which he did not

[1] See the very interesting paper of Thorndike on "Empirical Studies in the Theory of Measurement," *Archives of Psychology*, Vol. III. 1907. Also, H. Bruns, *Wahrscheinlichkeitsrechnung und Kollektivmasslehre*, 1906; and G. F. Lipps, *Die Theorie der Kollektivgegenstände*, 1902.

think before; or he may now see reason to take a more serious view of some factor previously tolerated under the belief of its harmlessness; or he may be facing some factor whose gravity he has all along realized well enough, but whose presence he has seen no way of escaping.

At this point, therefore, we have urgent need of some further statistical process, to enable us to estimate and eliminate such disturbing elements. And such has actually been discovered for us by Udny Yule. The fundamental significance of this event for the development of correlational research appears—both for the above reasons and for others of even greater importance—scarcely to admit of overestimation. The nature and usage of his corrective process have been fully explained elsewhere[1]. I need only mention here that it must be applied after eliminating the effect of the "accidents," not *vice versa*.

III. *Errors still persisting in the coefficient.*

Unfortunately, Yule's corrective process can rarely be carried out very completely; for though the calculations are simple enough, we almost always have great difficulty in rightly conceiving the nature of the chief irrelevant factors and also in obtaining sufficient information about them[2]. Hence it is useful to consider in what manner the coefficient will be affected by any errors still adhering to it.

$$\text{Let}\quad \bar{r}_{xy}=\frac{S(\bar{x}\bar{y})}{\sqrt{S(\bar{x}^2).S(\bar{y}^2)}}\ \frac{S(xy+xt+ys+st)}{\sqrt{S(x^2+2xs+s^2)\,S(y^2+2yt+t^2)}},$$

where $\bar{r}_{xy}$ denotes the coefficient obtained finally; s and t represent the errors in x and y that have still escaped elimination. Then $S(xt)$, $S(ys)$, $S(st)$, $S(xs)$, $S(s^2)$, $S(yt)$ and $S(t^2)$ are the values still possibly disturbing the coefficient. The above expression serves to show whether the disturbance tends to make the coefficient too large or too small.

We will take some cases that might occur in psychological experiments. $S(xt)$ would retain an appreciable value when there had been imperfect elimination of any correlation between x and the bias in the

[1] See Yule, *Proc. R. S. London*, Vol. LX. Also my own paper, *Am. Journ. Psychology*, Vol. XVIII. p. 161.

[2] This raises a vital question. At what point can the corrective process be considered as perfectly complete? This point is often not reached, I believe myself, until the corrected coefficient becomes = complete unity, or else zero. Here, and here only, the law under investigation has been completely disentangled from all other interfering factors. This consideration should, I think, dominate correlational research. In it is revealed the extreme significance of Yule's formula.

measurement of y. This would occur, for example, if the individuals who had a weak ability for the performance x were allowed to notice their inferiority, and thereby gained an extra stimulus to do well in y: the effect would be to make the final coefficient too large. If the individuals were not stimulated but only depressed by their insuccess in x, $S(xt)$ would take a negative value, and the final coefficient would be too small. On the whole, with moderately good experimentation, this source of error should generally be negligible. $S(yt)$ is, of course, similar.

$S(xs)$, or $S(yt)$, becomes appreciable when there is an uneliminated correlation between an ability and the bias in its measurement. This might occur, for instance, when the method of marking tended to exaggerate the differences between the good performances as compared with the bad ones. Its effect is to make the final coefficient too small. It can be avoided by basing the calculation on ranks instead of on measurements. In any case it does not seem formidable.

More serious is the liability to appreciable values of $S(st)$. This occurs when the bias in the measurement of x is correlated with that as regards y. For instance, if two performances are tested always in the morning, the measurements of those individuals who cannot do their best until the evening will be unfavourably biassed in both performances. Such a correlation, it may be noticed, will almost always be positive and there tend to make the final correlation too large.

But by far the greatest danger lies in $S(s^2)$, or $S(t^2)$. For the errors entering into these sums, being squared, become positive and thereby *lose all tendency to neutralize one another on being added up.* To reduce this danger is, in fact, the purpose of the correlational formula proposed in this paper. And this formula will be quite effectual in so far as two measurements can be obtained of x, such that the errors in the one are really independent of those in the other. But often it will be found impossible to avoid the same bias pervading more or less both measurements. A notable instance is when the "intelligence" of a class of children is estimated by two different teachers. Under ordinary circumstances, it is found that the two estimates show high correspondence, about ·80 or more. But when it can be arranged that the two teachers really form their estimates independently of one another, do not discuss the children together, nor hear of the same examination results, etc., then this correspondence shows a surprising shrinkage, thus revealing the previous high coefficient to have been spurious. In this way, however, the coefficient given by the proposed formula can only become too small; that is to say, the correction introduced by the formula is quite right as far as it goes: its only fault is in not going far enough, owing to the defectiveness of the data.

On the whole, it is clear that to obtain an approximation to the true correlational coefficient is by no means a simple matter. My own experience leads me to think that the sources of error considered above are more insidious even than the error arising from taking a small sample. The very large and imposing series of cases, which have been obtained at the expense of all the other moments of accuracy, are but as "whited sepulchres."

IV. *Amount of increased reliability to be obtained by increasing the number of measurements.*

A very convenient conception is that of the "reliability coefficient" of any system of measurements for any character. By this is meant the coefficient between one half and the other half of several measurements of the same thing, the division of the measurements into two halves or groups being done as described on p. 274.

It is often very useful to be able to estimate how much this reliability coefficient will probably be increased by any given additional number of measurements, or how much it will probably be reduced by any given diminution in the number of measurements. It can be shown that the following relation holds good:

$$\mathbf{r}_{x[p],x[p]} = \frac{\mathbf{p} \cdot \mathbf{r}_{x[q],x[q]}}{\mathbf{q} + (\mathbf{p} - \mathbf{q})\mathbf{r}_{x[q],x[q]}} \quad \ldots\ldots\ldots\ldots\ldots \text{(III)}^1,$$

where $r_{x[q],x[q]}$ is the known reliability coefficient of x when the latter has been measured $2q \times i$ times, i being any number, and $r_{x[p],x[p]}$ is the required most probable reliability coefficient if x be measured $2p \times i$ times.

Here, and also in the following section, all the measurements of x (or of y) are supposed to have been of equal general accuracy. But the case of unequal accuracy also admits of solution, see appendix *b*.

V. *Increase of correlation between two different characters to be obtained by increasing the number of measurements.*

The above principle can be usefully applied to the correlation between two different characters.

[1] For proof, see appendix *b*.

It will be found that[1]

$$r_{xy} = \frac{2\,(\mathrm{p}-1)\,(\mathrm{q}-1)\,\mathrm{r}_{x[1],y[1]} \times \mathrm{r}_{x[p],y[q]}}{2\sqrt{\mathrm{pq}\,(\mathrm{p}-1)(\mathrm{q}-1)}\,\mathrm{r}_{x[1],y[1]} - (\mathrm{p}+\mathrm{q}-2)\,\mathrm{r}_{x[p],y[q]}} \quad ..(\mathrm{IV}),$$

where $r_{x[1],y[1]}$ denotes the correlation of the average of any number, say i, measurements of x with that of any number, say u, of y; $r_{x[p],y[q]}$ denotes the correlation of the average of $p \times i$ measurements of x with that of $q \times u$ of y, and r_{xy} is, as before, the correct correlation between x and y.

This formula may especially be of service in supplementing the previous method, equation (I). For any inadequate fulfilment of the assumptions of both methods will, in general, affect their respective results differently. Hence, concordance of the results by both methods will greatly strengthen the evidence for the assumptions being valid.

An empirical formula was previously given by me for the same purpose (*Am. Jour. Psych.*, Vol. XV. pp. 88—91). Though of a different form from the above, it gives very similar results under the usual values of the terms entering into it. Still more accurately corresponding with the above theoretical formula are the empirical results given by Thorndike in his above-mentioned paper[2].

VI. *Discussion of some criticisms.*

Some time ago I had to take exception to the work of Professor Karl Pearson, on the ground that it is vitiated by observational errors and irrelevant factors[3]. To this he has made two replies[4]. But I am sorry to find that he has very seriously misrepresented my views, and even misquoted my figures. I can only regret all obscurities on my part that doubtlessly have contributed to this confusion, and hope that the following may elucidate matters.

My chief weakness he finds in what he takes to be the treatment of the "probable errors," about which he makes many strong comments. But here it is necessary to distinguish clearly between errors of two different sorts, those of sampling and those of observation. Suppose, for instance, that we wanted to investigate the head-length of skulls of prehistoric Thebes. We should only be able to obtain a limited number of them, and would have to assume that these were an adequate sample of the

[1] For proof, see appendix *c*. The formula simplifies greatly in the usual case that $p=q$.

[2] *Archives of Psychology*, Vol. III. p. 41, 1907.

[3] *Am. J. Psych.* Vol. XV. pp. 96—8.

[4] *Biometrika*, Vol. III. p. 160. *Drapers' Company Research Memoirs*, Biometric Series, IV. 1907.

whole; hereby an error is involved, whose general magnitude is measured by the "probable error" of sampling. But a totally new danger comes on the scene in the process of actually observing the head-lengths; if our instrument or method of observation is sufficiently crude, we may make here additional errors, even far exceeding those of sampling. Now, it is to these accidental errors of observation that my attention has been chiefly drawn. Pearson, on the other hand, while he has made widely and justly appreciated contributions to the theory of sampling errors, has scarcely touched on the errors of observation. And he seems to have remained "eingestellt" for sampling errors when reading my papers; for he takes in this sense everything that I really wrote about the errors of observation. Hence, of course, nothing but cross-purposes.

Take, to begin with, the main point at issue, correlation by ranks as compared with that by measurements. Suppose that there has been an examination in Latin and in Mathematics; we want to see how far a boy's success in the one subject agrees with that in the other. Evidently, there are two ways possible. We can note that he has got, say, 98 marks in Latin, but only 49 in Mathematics; this is called the method of measurements. Or we can remark that, out of 100 boys, he is first in Latin, but only 60th in Mathematics; this is the method of ranks. Under certain circumstances the method of ranks had seemed to me to be the less affected by the observational errors. This is especially the case where these errors increase in size towards one or both extremities of the range under consideration. In physical measurements, indeed, this will be rare; a large head is measured as easily as a small one, and in neither case should the error be excessive. But in psychology this is otherwise; the following, for instance, are the different thresholds for pitch found for 24 children, the unit being 1/3 v.d.: 5, 8, 10, 10, 13, 14, 14, 15, 17, 17, 18, 18, 20, 24, 25, 28, 33, 40, 45, 60, 70, 70, 70, 90. Any experienced psychologist will know that the error of determination at the bottom of this scale is at least some twenty times as great as at the top; to ignore this large inequality is to distribute "weight" very wrongfully, and therefore to do much injury to the reliability of the calculated coefficient. Translation from measurements into ranks is in such cases equivalent to a readjustment of the "weights" so as to equalize the upper and lower halves of the scale and reduce the importance of the extremities; no doubt, such device admits of much improvement. Now, Pearson takes all this about the observational error to refer to the error of sampling, expresses his disapproval most emphatically, and thinks to overthrow it by demonstrating that ranks and measurements produce, on the assumption of Gaussian distribution, *sampling* errors of quite equal magnitude. This, of course, has nothing to do with the point[1].

Let us take the next most important matter, the question of "squaring the differences." In our above example, the most natural way of estimating the degree of correspondence between the two examinations would be to notice the differences between the results of the one and those of the other (in doing so, we might either regard the differences of marks or those of rank). But the Bravais method of "product moments" introduces a refinement; it bases itself, not on these differences simply, but on their squares. I suggested in this Journal (Vol. II. Part 8, 1906) that under certain circumstances the omission of the squaring might reduce the accidental error (using this expression as equivalent to the observational error plus sampling error). I gave a formula for that purpose, which we will here term the R formula. In answer to this Pearson demonstrated that the sampling error by the R method is not less, but greater than that by the Bravais or r method, which he uses. But, as before, his demonstration does not touch the error of observation. Further, it is based on the assumption of a Gaussian distribution; and this assumption is a most precarious one, especially as regards the latter kind of error. It could easily be shown that there are other distributions where, on the contrary, the squaring is disadvantageous, just as, under similar circumstances, the average becomes less reliable than the median. Müller, Kraepelin, and others have shown such distributions to occur largely in psychological work, and my experience has often led me to suspect their influence in correlations also. It must be remembered that squaring lays stress on the extreme discrepancies between the series compared (not, as some people have said, on the extreme values in the series); and the reliability of these is often gravely in question. To take an example, suppose that all the individuals with one single large exception have shown a close relation between their performances in one experiment and in another; is there no ground for fearing that the one exception may be due to some accident, such as misunderstanding the procedure required, etc.? Only in this way can I explain that

[1] As regards the general question of ranks and measurements, it is pleasing to find that Udny Yule totally disagrees with Pearson's adverse comments and, on the contrary, finds my proposal of ranks "a very important step in the simplification of methods dealing with non-measurable character" (*Stat. Soc. Journ.* Vol. LXX. 1907, p. 656).

I have often found successive samples from the same class of events to fluctuate less when calculated by R than by r. The fact is that the Gaussian assumption is only a mathematical make-shift; we may often conveniently enough reckon formulae from it; but in actual application, we should constantly bear in mind its real limitations.

Seeing that, at any rate on the Gaussian assumption, the R method has a slightly larger probable error, Pearson severely criticizes the fact that I had found this to be only about $\cdot 43\sqrt{n}$, whereas that of r shows a much larger figure, $\cdot 67\sqrt{n-1}$. To explain this "paradox" he points out that, while the R method is only applicable to positive correlations, its probable error is taken from both positive and negative ones; and that the latter have a smaller range than the former. But there is no real paradox at all. We can no more argue that a probable error of ·43 by one method is smaller than a probable error of ·67 by another method, than we can say that 5 pounds are less than 10 francs. Two different methods, as a rule, are expressed in terms of different value; before making any such comparisons we must reduce them to common terms. And on doing so, the apparent superiority of R, of course, disappears. As regards the discrepancy emphasized by Pearson between the extreme positive and the extreme negative values, this seems to be of minor importance; he overlooks my empirical evidence that the *mean* positive and negative variations are nearly equal; and this is now corroborated by his own tables, as these show that the asymmetry between the positive and negative values only becomes marked in the extreme ranges, where the frequency—and therefore the effect on the probable error—is very small.

More serious is the charge against the R method, that it cannot deal directly with large negative correlations, but has first to convert these into positive ones by inverting one of the orders compared; and sometimes this is impossible, for even after the inversion the correlation may still remain negative. On looking into this more closely, however, its formidable appearance greatly diminishes; for it occurs solely when the correlation is so small as practically to be equivalent to zero, and therefore has no need of inversion at all; even in the extreme case selected by Pearson, the negative correlation is less than its own probable error.

Still such an anomaly, however harmless in actual practice, does indeed, I must admit, disqualify the R method from setting up to be a perfectly independent method ranking equally with the r method, still less claiming a large superiority over it. But how Pearson ever came to conceive that I made such a claim I have failed to discover. Far from doing so, and despite the above-mentioned occasional superiority of the R, I expressly entitled it a "footrule," as lying half-way between the r method with its complications (which I likened to an "elaborate micrometer") and judgment without mathematical method at all (which I compared to a "mere glance of the eye"). It is hard to understand how such strong expressions as these should ever have been taken to mean just the contrary. R's chief mission is merely to gain quickly an approximate valuation of r. As an example of the kind of work for which it was intended, I had occasion to put some 50 persons through a number of tests, principally as a demonstration; the work was rough, but still not so bad as to prevent all interest in the results. As there were 270 correlations to calculate, I could not possibly have attempted the task but for the extreme facility of R. Further, the method seemed well adapted for the schoolmaster who wants to know how far this year's examination tallies with that of last year, how far success in one subject has gone with that in another, how closely two teachers agree with one another in their estimates of children, and many more such problems. Seeing that R is meant to be subsidiary to r, the only real question is whether or not it actually produces values sufficiently approximating to the latter. Pearson selects a number of cases to prove discrepancy between the two; but he overlooks the fact that the discrepancy is never more than double the probable error, and that he himself declares any result less than 2—3 times the probable error to be devoid of significance[1]. If we take a general impartial review of the evidence hitherto adduced (for instance, that of Burt, Wimms, and Brown[2]), the correspondence of R with r appears to be amply good enough for the purpose in view; in fact R seems quite usable, not merely for assay purposes as originally contemplated, but even sometimes for research.

And even such discrepancies as do occur between the two coefficients are by no means wholly chargeable to the fault of R, as Pearson assumes[3]. For the differences between the two are only that R uses ranks and omits squaring; and both these differences, as we have seen, are often advantageous, so that then the discrepancies are more the fault of r. It may seem contradictory that R should under any circumstances

[1] For instance, *Drapers' Company Research Memoirs*, Biom. Series, IV. p. 15.

[2] Burt, *B. J. P.*, Vol. IV. p. 107. Wimms communicated his data to the Brit. Psych. Soc., Jan. 1910. For Brown, who is much the fullest on this point, see his paper quoted, p. 357.

[3] *Ibidem*, p. 38.

claim to be more accurate than its own ideal, r. But we must remember that its ideal is the true r, not the actually calculated one; the latter is the true one plus various errors (sampling, observation, irrelevant factors, etc.).

Pearson remarks repeatedly, and even italicizes, that I state the probable error of R to be $\cdot 4266/\sqrt{n}$, instead of $\cdot 4266/\sqrt{n-1}$. It is gratifying that criticism should have to turn so much to such a trivial matter. As it happened, however, I really made neither statement, but the quite accurate one, that the probable error is "$\cdot 43/\sqrt{n}$ with two correct decimals, when n is not less than 10[1]."

In spite of his attack on ranks, he has made an interesting contribution towards their use. He has worked out the relation between the coefficients of ranks (or rather "grades") and those of measurements, assuming Gaussian distribution. It is $r = 2\sin\left(\frac{\pi}{6}\rho\right)$, where ρ denotes the coefficient for ranks. On calculation, r and ρ turn out to be almost identical, thus corroborating my empirical observation to the same effect.

He has also done us the service of demonstrating, on the Gaussian assumption, the relation between squaring and non-squaring. It is

$$\sin\frac{\pi}{6}\rho = \cos\frac{\pi}{3}(1 - R) - 1.$$

I had found by actual observation the empirical formula

$$\rho = \sin\left(\frac{\pi}{2}R\right).$$

The theoretical values, it will be found, fit the observational ones with admirable closeness, the mean discrepancy being under $\cdot 01$ and the maximum only about $\cdot 02$, amounts that are negligible, at any rate in psychology. Hence, it is no small surprise that Pearson several times reproaches the empirical formula with being "erroneous." Even had there been any significant discrepancy between the two, it would not have affected the validity of the empirical value expressing actual observations, but only of the theoretical one based on such a weak assumption. And as, on the contrary, the discrepancy is so completely insignificant, there appears no great advantage, even on the Gaussian assumption, in abandoning the use of the older and simpler formula for ordinary rough purposes.

[1] *Brit. J. Psych.* Vol. II. p. 108.

On the whole, if we eliminate all these misapprehensions and oversights, there seems to be no serious difference of opinion on all these points between Pearson and myself. And to judge from the continually rising importance attributed by his school to observational errors and to irrelevant factors, even here the gap between us would appear to be rapidly closing.

Appendix.

a. Coefficient in the case of 2 *groups of measurements.*

Let $x(f_k)$ denote the kth measurement of the fth individual.

Let the superposed "accidental" disturbance be denoted by $d(f_k)$; the underlying "regular" measurement by $x'(f_k)$; the average value of $x'(f_k)$ by $x(f)$; and the kth "regular" deviation from this average by $u(f_k)$.

Then
$$x(f_k) = x'(f_k) + d(f_k) = x(f) + u(f_k) + d(f_k).$$

Let the measurements of each individual be divided into two groups, say a and b, in such a manner that any discrepancies between the averages of the two groups may be regarded as quite "accidental" (see p. 274).

Let the average of the values of $x(f_k)$, $x'(f_k)$, $d(f_k)$ that occur in group a be denoted by $x_a(f)$, $x_a'(f)$, $d_a(f)$. And let the average of *all* the values of $x(f_k)$, $x'(f_k)$, $u(f_k)$ be denoted by $x_{ab}(f)$, $x_{ab}'(f)$, $u_{ab}(f)$.

Then $\quad x_a(f) = x_a'(f) + d_a(f) = x_{ab}'(f) + d_a(f),$

since by assumption $x_a'(f) = x_b'(f) = x_{ab}'(f),$

$$= x(f) + u_{ab}(f) + d_a(f)$$
$$= x(f) + d_a(f), \text{ since } u_{ab}(f) = 0.$$

And $\quad x_{ab}(f) = x(f) + u_{ab}(f) + d_{ab}(f) = x(f) + d_{ab}(f).$

Analogously
$$x_b(f) = x(f) + d_b(f),$$
$$y_a(f) = y(f) + e_a(f),$$
$$y_b(f) = y(f) + e_b(f),$$
$$y_{ab}(f) = y(f) + e_{ab}(f).$$

Hence, summing for all individuals,

$$\frac{S(x_{ab}y_{ab})}{\sqrt{S(x_a x_b) \cdot S(y_a y_b)}} = \frac{S(x + d_{ab})(y + e_{ab})}{\sqrt{S(x + d_a)(x + d_b) \cdot S(y + e_a)(y + e_b)}}$$

$$= \frac{S(xy + xe_{ab} + yd_{ab} + d_{ab}e_{ab})}{\sqrt{S(x^2 + xd_a + xd_b + d_a d_b) \cdot S(y^2 + ye_a + ye_b + e_a e_b)}}$$

$$= \frac{S(xy)}{\sqrt{S(x^2) \cdot S(y^2)}} = r_{xy} \quad\ldots\ldots\ldots(1),$$

since each of the sums $S(xe)$, $S(yd)$, $S(de)$, etc. $= 0$, as we will assume, in accordance with pp. 273–4, that the d's and the e's are uncorrelated with one another and with x and y.

Further, $S(d_a^2)$ and $S(d_b^2)$ will be assumed to have not very dissimilar magnitudes (see p. 274). Hence *a fortiori* $S(x_a^2)$ and $S(x_b^2)$ will not be very unequal, so that approximately

$$\sqrt{S(x_a^2) \cdot S(x_b^2)} = \tfrac{1}{2}[S(x_a^2) + S(x_b^2)] \quad\ldots\ldots\ldots(2).$$

Hence
$$1 + \frac{1}{r_{x_a x_b}} = \frac{S(x_a x_b) + \sqrt{S(x_a^2) \cdot S(x_b^2)}}{S(x_a x_b)}$$

$$= \frac{2S(x_a x_b) + S(x_a^2) + S(x_b^2)}{2S(x_a x_b)}$$

$$= \frac{2S(x_{ab}^2)}{S(x_a x_b)} \quad\ldots\ldots\ldots(3).$$

And, similarly,
$$1 + \frac{1}{r_{x_a x_b}} = \frac{2S(y_{ab}^2)}{S(y_a y_b)} \quad\ldots\ldots\ldots(4).$$

Then, as

$$\frac{S(x_{ab}y_{ab})}{\sqrt{S(x_a x_b) \cdot S(y_a y_b)}} = \frac{S(x_{ab}y_{ab})}{\sqrt{S(x_{ab}^2) \cdot S(y_{ab}^2)}} \cdot \sqrt{\frac{S(x_{ab}^2) \cdot S(y_{ab}^2)}{S(x_a x_b) \cdot S(y_a y_b)}},$$

we get by (1), (3) and (4)

$$r_{xy} = r_{x_{ab}y_{ab}} \cdot \frac{1}{2}\sqrt{\left(1 + \frac{1}{r_{x_a x_b}}\right)\left(1 + \frac{1}{r_{y_a y_b}}\right)} \quad\ldots\ldots\ldots(5).$$

b. Proof of formula III.

Take now the more general case of p, instead of 2, groups of measurements for x, denoted by

$$x_1,\ x_2, \ldots x_k \ldots x_h \ldots x_p = x + d_1,\ x + d_2, \ldots x + d_p,$$

where x is the underlying regular measurement, while the d's are the superposed accidental components.

Let $\frac{x_1+x_2+\ldots+x_p}{p}$ be denoted by $x[p]$, $\frac{d_1+d_2+\ldots+d_p}{p}$ by $d[p]$.

Let $y_1, y_2, \ldots y_u \ldots y_v \ldots y_s, e_1, e_2, \ldots e_s, y[s], e[s]$ have similar meanings with regard to y.

Since, as we have seen, such sums as $S(xe)$, $S(xd)$, $S(d_k d_h)$ each $=0$, we get, summing for all individuals,

$$S(x[p]\,.\,y[s]) = S(x+d[p])(y+e[s]) = S(xy),$$

$$S(x_k x_h) = S(x+d_k)(x+d_h) = S(x^2),$$

and
$$S(y_u y_v) = S(y+e_u)(y+e_v) = S(y^2).$$

So that

$$r_{xy} = \frac{S(xy)}{\sqrt{S(x^2)\,.\,S(y^2)}} = \frac{S(x[p]\,.\,y[s])}{\sqrt{S(x^2[p])\,.\,S(y^2[s])}} \cdot \sqrt{\frac{S(x^2[p])\,.\,S(y^2[s])}{S(x^2)\,.\,S(y^2)}}$$

$$= r_{x[p],\,y[s]} \cdot \sqrt{\frac{\frac{S(x_1+x_2+\ldots+x_p)^2}{p^2}}{S(x_k x_h)}} \cdot \sqrt{\frac{\frac{S(y_1+y_2+\ldots+y_s)^2}{s^2}}{S(y_u y_v)}}$$

$$= r_{x[p],\,y[s]} \cdot \sqrt{\frac{S_k S(x_k^2) + p(p-1)\,S(x_k x_h)}{p^2 S(x_k x_h)}}$$

$$\times \sqrt{\frac{S_u S(y_u^2) + s(s-1)\,S(y_u y_v)}{s^2 S(y_u y_v)}} \quad \ldots\ldots(6),$$

the additional S denoting summation for all groups.

But, from (2),

$$\frac{pS(x_k x_h)}{S_k S(x_k^2)} = \frac{2}{p(p-1)} S_{kh} \left[\frac{S(x_k x_h)}{\sqrt{S(x_k^2)\,.\,S(x_h^2)}}\right] = \frac{2}{p(p-1)} S_{kh}(r_{x_k x_h})$$

$=$ average correlation between x_k and x_h = say, $r_{x[1],\,x[1]} \ldots (7)$.

From (6) and (7), putting $y=x$, we get

$$1 = r_{x[p],\,x[s]} \cdot \sqrt{\frac{1}{p r_{x[1],\,x[1]}} + \frac{p-1}{p}} \cdot \sqrt{\frac{1}{s r_{x[1],\,x[1]}} + \frac{s-1}{s}},$$

from which

$$r_{x[p],\,x[s]} = \sqrt{\frac{p r_{x[1],\,x[1]}}{1+(p-1)\,r_{x[1],\,x[1]}}} \cdot \sqrt{\frac{s r_{x[1],\,x[1]}}{1+(s-1)\,r_{x[1],\,x[1]}}} \ldots (8).$$

In the usual case that $s=p$, this becomes

$$r_{x[p],\,x[p]} = \frac{p r_{x[1],\,x[1]}}{1+(p-1)\,r_{x[1],\,x[1]}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(9),$$

or, writing q for p,

$$r_{x[q],\,x[q]} = \frac{q r_{x[1],\,x[1]}}{1+(q-1)\,r_{x[1],\,x[1]}}.$$

And the two last equations give, on reduction,

$$r_{x[p],\,x[p]} = \frac{p r_{x[q],\,x[q]}}{q + (p-q)\, r_{x[q],\,x[q]}} \qquad \text{..................(10).}$$

Although this formula applies immediately to groups of approximately equal liability to accidental disturbances, it can easily be extended to cases of unequal liability. For an actual measurement of any degree of accuracy is, clearly, equivalent to the average of a number of measurements of an inferior degree of accuracy. So that two actual measurements (or groups of such) of unequal accuracy may be conceived as the averages of two unequal numbers of measurements all of equal (inferior) accuracy. Thus p could represent $S_k(m_k g_k)$, where m indicates the number of groups, and g their respective precisions.

c. Proof of formulae I and IV.

In (5) let each of the groups, a and b, be composed of $\frac{p}{2}$ sub-groups each satisfying all the assumptions we made about a and b.

Then (5) may be written as

$$r_{xy} = r_{x[p],\,y[q]} \frac{1}{2} \sqrt{\left(1 + \frac{1}{r_{x\left[\frac{p}{2}\right],\,x\left[\frac{p}{2}\right]}}\right)\left(1 + \frac{1}{r_{y\left[\frac{q}{2}\right],\,y\left[\frac{q}{2}\right]}}\right)},$$

where the indices in brackets have the same signification as in appendix b.

This, owing to (8), becomes

$$= r_{x[p],\,y[q]} \cdot \sqrt{\frac{1+(p-1)\, r_{x[1],\,x[1]}}{p r_{x[1],\,x[1]}} \cdot \frac{1+(q-1)\, r_{y[1],\,y[1]}}{q r_{y[1],\,y[1]}}} \quad \text{...(11)}$$

$$= r_{x[p],\,y[q]} \cdot \sqrt{\frac{(p-1)(q-1)}{pq r_{x[1],\,x[1]} \cdot r_{y[1],\,y[1]}}} \cdot \sqrt{\frac{1}{p-1} + r_{x[1],\,x[1]}} \times \sqrt{\frac{1}{q-1} + r_{y[1],\,y[1]}}\,.$$

The two factors on the right

$$= \frac{1}{2}\left[r_{x[1],\,x[1]} + r_{y[1],\,y[1]} + \frac{1}{p-1} + \frac{1}{q-1} - \left(\sqrt{r_{x[1],\,x[1]} + \frac{1}{p-1}} - \sqrt{r_{y[1],\,y[1]} + \frac{1}{q-1}}\right)^2 \right]$$

$$= \sqrt{r_{x[1],\,x[1]} \cdot r_{y[1],\,y[1]}} + \frac{p+q+2}{2(p-1)(q-1)} + \tfrac{1}{2}\left(\sqrt{r_{x[1],\,x[1]}} - \sqrt{r_{y[1],\,y[1]}}\right)^2 - R, \text{ say......(12).}$$

Hence, from (11) (putting there $p=q=1$) and (12),

$$r_{xy}=r_{x[p],\,y[q]}\left(\sqrt{\frac{(p-1)(q-1)}{pq}}+\frac{p+q-2}{2\sqrt{pq(p-1)(q-1)}}\cdot\frac{r_{xy}}{r_{x[1],\,y[1]}}\right)+R', \text{ say}\ldots\ldots(13).$$

We will neglect R', which vanishes when the two coefficients ($r_{x[1],\,x[1]}$ and $r_{y[1],\,y[1]}$) and also the two numbers of groups (p and q) tend to equality; it becomes largest, and then positive, when the coefficient and also the number of groups for one character compared are both much greater than those for the other character.

We get then, finally, on reduction,

$$r_{xy}=\frac{2(p-1)(q-1)\,r_{x[1],\,y[1]}\cdot r_{x[p],\,y[q]}}{2\sqrt{pq(p-1)(q-1)}\,r_{x[1],\,y[1]}-(p+q-2)\,r_{x[p],\,y[q]}}\ldots(14).$$

d. Proof of formula II.

Let z denote

$$r_{12}\sqrt{\frac{1+(p-1)\,r_{34}}{pr_{34}}\cdot\frac{1+(q-1)\,r_{56}}{qr_{56}}},$$

where 1, 2, 3, 4, 5, 6 indicate any values.

Taking logs,

$$\log z=\log r_{12}+\tfrac{1}{2}\log\left[1+(p-1)\,r_{34}\right]-\tfrac{1}{2}\log pr_{34}+\tfrac{1}{2}\log\left[1+(q-1)\,r_{56}\right]-\tfrac{1}{2}\log r_{56},$$

and differentiating

$$\frac{dz}{z}=\frac{dr_{12}}{r_{12}}-\frac{1}{2}\,\frac{dr_{34}}{r_{34}\left[1+(p-1)\,r_{34}\right]}-\frac{1}{2}\,\frac{dr_{56}}{r_{56}\left[1+(q-1)\,r_{56}\right]}.$$

Square all such equations, add, and divide by the number of equations, then

$$\begin{aligned}\frac{\sigma_z^2}{z}=\frac{\sigma_{r_{12}}}{r_{12}^2}&+\frac{\sigma_{r_{34}}^{\;2}}{4r_{34}^2\left[1+(p-1)\,r_{34}\right]^2}\\&+\frac{\sigma_{r_{56}}^{\;2}}{4r_{56}^2\left[1+(q-1)\,r_{56}\right]^2}-\frac{\sigma_{r_{12}}\sigma_{r_{34}}R_{r_{12}r_{34}}}{r_{12}r_{34}\left[1+(p-1)\,r_{34}\right]}\\&-\frac{\sigma_{r_{12}}\sigma_{r_{56}}R_{r_{12}r_{56}}}{r_{12}r_{56}\left[1+(q-1)\,r_{56}\right]}+\frac{\sigma_{r_{34}}\sigma_{r_{56}}R_{r_{34}r_{56}}}{2r_{34}r_{56}\left[1+(p-1)\,r_{34}\right]\left[1+(q-1)\,r_{56}\right]},\end{aligned}$$

where[1]

$$R_{r_{12}r_{34}}=\begin{bmatrix}(r_{13}-r_{12}r_{23})(r_{24}-r_{23}r_{34})\\+(r_{14}-r_{34}r_{13})(r_{23}-r_{12}r_{13})\\+(r_{13}-r_{14}r_{34})(r_{24}-r_{12}r_{14})\\+(r_{14}-r_{24}r_{12})(r_{23}-r_{24}r_{34})\end{bmatrix}\div 2(1-r_{12}^2)(1-r_{34}^2),$$

with similar expressions for $R_{r_{34}r_{56}}$ and $R_{r_{12}r_{56}}$.

[1] See Pearson and Filon, *Phil. Trans.* A, Vol. CXCI. p. 262. I am greatly obliged to Professor Filon, not only for his valuable paper, but also for being kind enough to send me equation (15) deduced by its means for the case that $p=q=2$. Above I give his deduction, but generalized to include all values of p and q.

Also, from the same paper, $\sigma_{r_{12}} = \dfrac{1 - r_{12}^2}{\sqrt{n}}$.

Thus we have

$$\frac{n\sigma_z^2}{z^2} = \frac{(1 - r_{12}^2)^2}{r_{12}^2} + \frac{(1 - r_{34}^2)^2}{4r_{34}^2[1 + (p-1)r_{34}]^2} + \frac{(1 - r_{56}^2)^2}{4r_{56}^2[1 + (q-1)r_{56}]^2}$$

$$+ \frac{1}{2r_{12}r_{34}[1 + (p-1)r_{34}]} \times \begin{bmatrix} (r_{13} - r_{12}r_{23})(r_{24} - r_{23}r_{34}) \\ + (r_{14} - r_{34}r_{13})(r_{23} - r_{12}r_{13}) \\ + (r_{13} - r_{14}r_{34})(r_{24} - r_{12}r_{14}) \\ + (r_{14} - r_{24}r_{12})(r_{23} - r_{24}r_{34}) \end{bmatrix}$$

$$+ \frac{1}{2r_{12}r_{56}[1 + (q-1)r_{56}]} \times \begin{bmatrix} (r_{15} - r_{12}r_{52})(r_{26} - r_{25}r_{65}) \\ + (r_{15} - r_{16}r_{56})(r_{26} - r_{21}r_{61}) \\ + (r_{16} - r_{15}r_{65})(r_{25} - r_{21}r_{51}) \\ + (r_{16} - r_{12}r_{62})(r_{25} - r_{26}r_{56}) \end{bmatrix}$$

$$+ \frac{1}{4r_{34}r_{56}[1 + (p-1)r_{34}][1 + (q-1)r_{56}]} \times \begin{bmatrix} (r_{35} - r_{34}r_{54})(r_{46} - r_{45}r_{65}) \\ + (r_{35} - r_{36}r_{56})(r_{46} - r_{43}r_{63}) \\ + (r_{36} - r_{34}r_{64})(r_{45} - r_{46}r_{56}) \\ + (r_{36} - r_{35}r_{65})(r_{45} - r_{43}r_{53}) \end{bmatrix} \quad \text{......(15).}$$

This equation (15) holds good for all values of the indices 1, 2, 3, 4, 5, 6. Let them now be replaced by $x[p]$, $y[q]$, x_k, x_h, y_u, y_v, where these terms have the same meaning as in appendix *b*.

Then, as the indices k and h indicate groups of measurements differing from one another only in the distribution of the accidental disturbances among the individuals, k and h may legitimately be conceived to have such values that, in general, the correlations produced both by x_k and by x_h are equal to the average of the correlations produced by all the groups for x. Analogously, as regards the indices u and v.

We get, then,

$$r_{12} = r_{x[p],\, y[q]} = \text{say}, f \quad \text{........................(16),}$$

$$r_{34} = r_{x[1],\, x[1]} = \text{say}, g \quad \text{........................(17),}$$

$$r_{56} = r_{y[1],\, y[1]} = \text{say}, h \quad \text{........................(18).}$$

Also from (8) $r_{13} = r_{14} = r_{x[p],\, x[1]} = \dfrac{\sqrt{p}\,.\,g}{\sqrt{1 + (p-1)\,g}}$ ....(19),

and $r_{25} = r_{26} = r_{y[q],\, y[1]} = \dfrac{\sqrt{q}\,.\,h}{\sqrt{1 + (q-1)\,h}}$ ....(20).

And $r_{35} = r_{45} = r_{36} = r_{46} = r_{x[1],\, y[1]}$, which, utilising (11),

$$= f \frac{\sqrt{[1+(p-1)g][1+(q-1)h]}}{\sqrt{pq}} \quad \text{...............(21)},$$

$$r_{23} = r_{24} = r_{x[1],\, y[q]} = f \frac{\sqrt{1+(p-1)g}}{\sqrt{p}} \quad \text{............(22)},$$

$$r_{15} = r_{16} = r_{x[p],\, y[1]} = f \frac{\sqrt{1+(q-1)h}}{\sqrt{q}} \quad \text{............(23)}.$$

Then, as $\qquad z = f \sqrt{\frac{1+(p-1)g}{pg} \cdot \frac{1+(q-1)h}{qh}}$,

substituting from (16)—(23) in equation (15), we get finally

$$\sigma_z^2 = \frac{1}{n} \frac{1+(p-1)g}{pg} \cdot \frac{1+(q-1)h}{qh} \left[(1-f^2)^2 + f^2 \frac{(1-g)[1+g-9g^2-(4p-9)g^3]}{4g^2[1+(p-1)g]^2} \right.$$
$$\left. + f^2 \frac{(1-h)[1+h-9h^2-(4q-9)h^3]}{4h^2[1+(q-1)h]^2} + f^4 \frac{qh(1-g)+pg(1-h)+(1-g)(1-h)}{pqgh} \right] \quad \text{......(24)}.$$

The three terms on the right are usually small. Neglecting them, we get with sufficient approximation for most psychological purposes,

$$\sigma_z = \frac{1-f^2}{\sqrt{n}} \sqrt{\frac{1+(p-1)g}{pg} \cdot \frac{1+(q-1)h}{qh}},$$

or the probable error of the correct coefficient r_{xy}

$$= \cdot 6745 \frac{1 - r^2_{x[p],\, y[q]}}{\sqrt{n}} \cdot \frac{r_{xy}}{r_{x[p],\, y[q]}},$$

where $r_{x[p],\, y[q]}$, as in (II), denotes the correlation between the average of all the p group averages for x and the average of all the g ones for y.

e. *Yule's proof of the correction formula.*

x_1 and y_1 are measures of x and y at a certain series of measurements.

x_2 „ y_2 „ „ „ another „ „

Let $\qquad x_1 = x + \delta_1, \; x_2 = x + \delta_2, \; y_1 = y + \epsilon_1, \; y_2 = y + \epsilon_2,$

all terms denoting deviations from means.

Then, if it is assumed that δ, ϵ, the errors of measurement, are uncorrelated with each other or with x or y,

$$\Sigma(x\delta) \text{ etc.} = 0, \quad \Sigma(x_1 y_1) = \Sigma(xy).$$

Hence $$r_{x_1y_1}\sigma_{x_1}\sigma_{y_1} = r_{xy}\sigma_x\sigma_y,$$

and similarly $$r_{x_2y_2}\sigma_{x_2}\sigma_{y_2} = r_{xy}\sigma_x\sigma_y,$$

$$r_{x_1y_2}\sigma_{x_1}\sigma_{y_2} = r_{xy}\sigma_x\sigma_y,$$

$$r_{x_2y_1}\sigma_{x_2}\sigma_{y_1} = r_{xy}\sigma_x\sigma_y,$$

or $$r_{xy}^4 = r_{x_1y_1}\cdot r_{x_2y_2}\cdot r_{x_1y_2}\cdot r_{x_2y_1}\frac{\sigma_{x_1}^2\sigma_{x_2}^2\sigma_{y_1}^2\sigma_{y_2}^2}{\sigma_x^4\sigma_y^4} \quad \text{.........(1).}$$

But also, since $\Sigma(x\delta) = 0$, $\Sigma x_1x_2 = \Sigma x^2$,

and $$r_{x_1x_2}\sigma_{x_1}\sigma_{x_2} = \sigma_x^2,$$

or $$\sigma_{x_1}\sigma_{x_2} = \frac{\sigma_x^2}{r_{x_1x_2}} \text{ and } \sigma_{y_1}\sigma_{y_2} = \frac{\sigma_y^2}{r_{y_1y_2}} \quad \text{.................(2).}$$

From (1) and (2)

$$r_{xy}^4 = r_{x_1y_1}r_{x_2y_2}r_{x_1y_2}r_{x_2y_1}$$