Spearman, Charles Edward
The proof and measurment of association between two things

THE PROOF AND MEASUREMENT OF

ASSOCIATION

BETWEEN TWO THINGS.

By: C. Spearman,

From Amer Jour of Psy Vol. 15-1904

THE PROOF AND MEASUREMENT OF ASSOCIATION BETWEEN TWO THINGS

By C. Spearman

CONTENTS.

INTRODUCTORY.

All knowledge - beyond that of bare isolated occurrence- deals with uniformities. Of the latter, some few have a claim to be considered absolute, such as mathematical implications and mechanical laws. But the vast majority are only <u>partial;</u> medicine does not teach that smallpox is inevitably escaped by vaccination, but that it is so generally; biology has not shown that all animals require organic food, but that nearly all do so; in daily life, a dark sky is no proof that it will rain, but merely a warning; even in morality, the sole categorical imperative alleged by Kant was the sinfulness of telling a lie, and few thinkers since have admitted so much as this to be valid universally. In psychology, more perhaps than in any other science, it is hard to find absolutely inflexible coincidences; occasionally, indeed, there appear uniformities sufficiently regular to be practically treated as laws, but infinitely the greater part of the observations hitherto recorded concern only more or less pronounced tendencies of one event or attribute to accompany another.

Under these circumstances, one might well have expected that the evidential evaluation and precise mensuration of tendencies had long been the subject of exhaustive investigation

THE PROOF AND MEASUREMENT OF ASSOCIATION BETWEEN TWO THINGS

By C. Spearman

CONTENTS.

INTRODUCTORY.

All knowledge - beyond that of bare isolated occurrence - deals with uniformities. Of the latter, some few have a claim to be considered absolute, such as mathematical implications and mechanical laws. But the vast majority are only partial; medicine does not teach that smallpox is inevitably escaped by vaccination, but that it is so generally; biology has not shown that all animals require organic food, but that nearly all do so; in daily life, a dark sky is no proof that it will rain, but merely a warning; even in morality, the sole categorical imperative alleged by Kant was the sinfulness of telling a lie, and few thinkers since have admitted so much as this to be valid universally. In psychology, more perhaps than in any other science, it is hard to find absolutely inflexible coincidences; occasionally, indeed, there appear uniformities sufficiently regular to be practically treated as laws, but infinitely the greater part of the observations hitherto recorded concern only more or less pronounced tendencies of one event or attribute to accompany another.

Under these circumstances, one might well have expected that the evidential evaluation and precise mensuration of tendencies had long been the subject of exhaustive investigation and now formed one of the earliest sections in a beginner's psychological course. Instead, we find only a general naive ignorance that there is anything about it requiring to be learnt. One after another, laborious series of experiments are executed and published with the purpose of demonstrating some connection between two events, wherein the otherwise learned psychologist reveals that his art of proving and measuring correspondence has not advanced beyond that of lay persons. The consequence has been that the significance of the experiments is not at all rightly understood, nor have any definite facts been elicited that may be either confirmed or refuted.

The present article is a commencement at attempting to remedy this deficiency of scientific correlation. With this view, it will be strictly confined to the needs of practical workers, and all theoretical mathematical demonstrations will be ommitted; it may, however, be said that the relations stated have already received a large amount of empirical verification. Great thanks are due from me to Professor Haussdorff and to Dr. G. Lipps, each of whom have supplied a useful theorem in polynomial probability; the former has also very kindly given valuable advice concerning the proof of the important formulae for elimination of "systematic deviations."

At the same time, and for the same reason, the meaning and working of the various formulae have been explained sufficiently, it is hoped, to render them readily usable even by those whose knowledge of mathematics is elementary. The fundamental procedure is accompanied by simple imaginary examples, while the more advanced parts are illustrated by cases that have actually occurred in my personal experience. For more abundant and positive exemplification, the reader is requested to refer to the under cited research,[1] which is entirely built upon the principles and mathematical relations here laid down.

In conclusion, the general value of the methodics recommended is emphasized by a brief criticism of the best correlational work hitherto made public, and also the important question is discussed as to the number of "cases" required for an experimental series.

Part 1.
ELEMENTARY CORRELATION AND "ACCIDENTAL DEVIATION."

1. Requirements of a Good Method of Correlation.
(a) Quantitative expression.

The most fundamental requisite is to be able to measure our observed correspondence by a plain numerical symbol. There

---

1. "General Intelligence" determined and measured, to appear in a subsequent number of this Journal.

and now formed one of the earliest sections in an 'experimental' psychological course. Instead, we find only a general naive ignorance that there is anything about it requiring to be learnt. One after another, laborious series of experiments are executed and published with the purpose of demonstrating some connection between two events, wherein the otherwise learned psychologist reveals that his art of proving and measuring correspondence has not advanced beyond that of lay persons. The consequence has been that the significance of the experiments is not at all rightly understood, nor have any definite facts been elicited that may be either confirmed or refute.

The present article is a commencement at attempting to remedy this deficiency of scientific correlation. With this view, it will be strictly confined to the needs of practical workers, and all theoretical mathematical demonstrations will be omitted; it may, however, be said that the relations stated have already received a large amount of empirical verification. Great thanks are due from me to Professor Haussdorff and to Dr. G. Lipps, each of whom have supplied a useful theorem in polynomial probability; the former has also very kindly given valuable advice concerning the proof of the important formulae for elimination of 'systematic deviations.'

At the same time, and for the same reason, the meaning and working of the various formulae have been explained sufficiently, it is hoped, to render them readily usable even by those whose knowledge of mathematics is elementary. The fundamental procedure is accompanied by simple imaginary examples, while the more advanced parts are illustrated by cases that have actually occurred in my personal experience. For more abundant and positive exemplification, the reader is requested to refer to the under cited research,[1] which is entirely built upon the principles and mathematical relations here laid down.

In conclusion, the general value of the methodics recommended is emphasized by a brief criticism of the best correlational work hitherto made public, and also the important question is discussed as to the number of 'cases' required for an experimental series.

Part I.

ELEMENTARY CORRELATION AND 'ACCIDENTAL DEVIATION.'

1. Requirements of a good Method of Correlation.

(a) Quantitative expression.

The most fundamental requisite is to be able to measure our observed correspondence by a plain numerical symbol. There

---

1. 'General Intelligence' determined and measured, to appear in a subsequent number of this Journal.

is no reason whatever to be satisfied either with vague generalities such as "large,""medium,""small," or, on the other hand, with complicated tables and compilations.

The first person to see the possibility of this immense advance seems to have been Galton, who, in 1886, writes: "the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long leg and conversely."[1] He then proceeds to devise the required symbol in such a way that it conveniently ranges from 1, for perfect correspondence, to 0 for entire independence, and on again to -1 for perfect correspondence inversely. By this means, correlations became comparable with other ones found either in different objects or by different observers; they were at last capable of leading to further conclusions, speculative and practical; in a word, they now assumed a scientific character.

Mathematically, it is clear that innumerable other systems of values are equally concievable, similarly ranging from 1 to 0. One such, for instance, has been worked out and extensively used by myself (see pp 15 ff). It therefore becomes necessary to discuss their relative merits.

(b) The significance of the quantity.

Galton's particular system is defined and most advantageously distinguished from all the others by the important property, that if any number of arms, for instance, be collected which are all any amount, $x\sigma_a$ above the mean, then the corresponding legs will average $rx\sigma_1$ above the mean (with a middle or "quartile" deviation(2) of $\sigma_1\sqrt{1-r^2}$) ;where $\sigma_a =$ the quartile variation of the arms, $\sigma_1 =$ that of the legs, and r is the measure of the correlation.

But another- theoretically far more valuable - property may conceivable attach to one among the possible systems of values expressing the correlation; this is, that a measure might be afforded of the hidden underlying cause of the variations. Suppose, for example, that A and B both derive their money from variable dividends and each gets 1/x th. of his total from some source common to both of them. Then evidently their respective incomes will have a certain tendency to rise and fall simultaneously; this correspondence will in any of the possible systems of values always be some function 1/x, but in only one of them will it actually be itself 6 1/x; in such a favored case, if A and B get, say,20% of their respective incomes from the common source, the correlation between these two imcomes will also show itself as 0.20; and conversely, if A's income happens to be found correlated with that of B by 0.20, them

---

1."Proceedings Royal Society of London", Vols XL and XLV.

2. Commonly, but misleadingly, termed the "probable error."

is no reason whatever to be satisfied either with vague qualities such as "large," "medium," "small," or, on the other hand, with complicated tables and compilations.

The first person to see the possibility of this immense advance seems to have been Galton, who, in 1888, writes: "the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long leg and conversely."[1] He then proceeds to devise the required symbol in such a way that it conveniently ranges from 1, for perfect correspondence, to 0 for entire independence, and on again to -1 for perfect correspondence inversely. By this means, correlations became comparable with other ones found either in different objects or by different observers; they were at last capable of leading to further conclusions, speculative and practical; in a word, they now assumed a scientific character.

Mathematically, it is clear that innumerable other systems of values are equally conceivable, similarly ranging from 1 to 0. One such, for instance, has been worked out and extensively used by myself (see pp. 15 ff). It therefore becomes necessary to discuss their relative merits.

(b) The significance of the quantity.

Galton's particular system is defined and most advantageously distinguished from all the others by the important property, that if any number of arms, for instance, be collected which are all any amount, $x\sigma_a$ above the mean, then the corresponding legs will average $rx\sigma_l$ above the mean (with a middle or "quartile' deviation[2] of $\sigma_l\sqrt{1-r^2}$); where σ_a = the quartile variation of the arms, σ_l = that of the legs, and r is the measure of the correlation.

But another - theoretically far more valuable - property may conceivable attach to one among the possible systems of values expressing the correlation; this is, that a measure might be afforded of the hidden underlying cause of the variations. Suppose, for example, that A and B both derive their money from variable dividends and each gets 1/x th. of his total from some source common to both of them. Then evidently their respective incomes will have certain tendency to rise and fall simultaneously; this correspondence will in any of the possible systems of values always be some function 1/x, but in only one of them will it actually be itself = 1/x; in such a favored case, if A and B get, say, 20% of their respective incomes from the common source, the correlation between these two incomes will also show itself as 0.20; and conversely, if A's income happens to be found correlated with that of B by 0.20, then

---

1. "Proceedings Royal Society of London", Vols XL and XLV.

2. Commonly, but misleadingly, termed the "probable error."

there is a likelihood that 0.20 of A's income coincides with 0.20 of B, leaving to either 0.80 disposable independently. The observed correlation thus becomes the direct expression of the relative amount of underlying influences tending for and against the correspondence.

In the above imagined instance, this desirable expressiveness belongs to the same above system of values proposed by Galton (and elabotated by Pearson). But this instance is exceptional and fundamentally different from the normal type. Evidently, A and B need not necessarily derive exactly the same proportion of their incomes from the common sourse; A might get his 0.20 while B got some totally different share; in which case, it will be found that the correlation is always the geometrical mean between the two shares. Let B be induced to put all his income into the common fund, then A need only put in $0.20^2 = 0.04$, to maintain the same correlation as before; since the geometrical mean between 0.04 and 1 is equal to 0.20.

Now, in psychological, as in most other actual correspondences, A and B are not to be regarded as in the fixed bisection of our first case, but rather as in the labile inter-accommodation of our second case. Hence A, in order to be correlated with B by $1/x$, must be considered to have only devoted $1/x^2$ (instead of $1/x$) of his arrangement to this purpose and therefore still have for further arrangement $1 - 1/x^2$, which will enable an independent correlation to arise of $\sqrt{1-1/x^2}$. In short, not Galton's measure of correlation, but the <u>square thereof</u>, indicates the relative influence of the factors in A tending towards any observed correspondence as compared with the remaining components of A tending in other directions.

(c) Accuracy.

From this plurality of possible systems of values for the measure of the correlation must be carefully distinguished the variety of ways of calculating any one of them. These latter again, have various advantages and disadvantages, of which the principal is their respective degrees of liability to "accidental deviation."

For, though the correlation between two series of data is an absolute mathematical fact, yet its whole real value lies in our being able to assume a likelihood of further cases taking a similar direction; we want to consider our results as a truly representative sample. Any one at all accustomed to original investigation must be aware how frequently phenomena will group themselves in such a manner as to convincingly suggest the existence of some law - when still more prolonged experiment reveals that the observed uniformity was due to pure hazard and has no tendency whatever to further repeat itself.

Luckily, this one great source of fallacy can be adequately there is a likelihood that 0.10 of his income coincides with 0.20 of B, leaving to either 0.? disposable independently. The observed correlation thus becomes the direct expression of the relative amount of underlying influences tending for and against the correspondence.

In the above imaginary instance, this method extracts likeness belonging to the same above system of values proposed by Galton (and elaborated by Pearson). But this instance is exceptional and fundamentally different from the normal type. Evidently, A and B need not necessarily derive exactly the same proportion of their incomes from the common fund; A might get his 0.20 while B got some totally different share; in which case, it will be found that the correlation is always the geometrical mean between the two shares. Let B be reduced to put all his income into the common fund, then A need only put in $0.20^2 = 0.04$, to maintain the same correlation as before: since the geometrical mean between 0.04 and 1 is equal to 0.20.

Now, in psychological, as in most other actual correspondences, A and B are not to be regarded as in the fixed bisection of our first case, but rather as in the labile inter-accommodation of our second case. Hence, in order to be correlated with B by $1/x$, must be considered to have only devoted $1/x^2$ (instead of $1/x$) of his arrangements to this purpose and therefore still have for further arrangement $1 - 1/x^2$, which will enable an independent correlation to arise of $\sqrt{1-1/x^2}$. In short, not Galton's measure of correlation, but the square thereof, indicates the relative influence of the factors in A tending towards any observed correspondence as compared with the remaining components of A tending in other directions.

(c) Accuracy.

From this plurality of possible systems of values for the measure of the correlation must be carefully distinguished the variety of ways of calculating any one of them. These latter again, have various advantages and disadvantages, of which the principal is their respective degrees of liability to "accidental deviation."

For, though the correlation between two series of data is an absolute mathematical fact, yet its whole real value lies in our being able to assume a likelihood of further cases taking a similar direction; we want to consider our results as a truly representative sample. Any one at all accustomed to original investigation must have seen how frequently phenomena will group themselves in such a manner as to suggest prima facie the existence of some law - when still more prolonged experiment reveals that the observed uniformity was due to pure hazard and has no tendency whatever to further present itself.

Luckily, this one great source of fallacy can be adequately eliminated, owing to the fact that such accidental deviations are different in every individual case (hence are often called the "variable errors") and occur quite impartially in every direction according to the known laws of probability. The consequence is that they eventually more or less completely <u>compensate one another</u>, and thus finally present an approximately true result. Such elimination, however, must always remain theoretically imcomplete, since no amount of chance coincidence is absolutely impossible; but beyond certain limits it becomes so extremely unlikely that for practical purposes we can afford to neglect it. When a person loses 14 times running at pitch-and-toss, he can reckon that such a series would not occur by mere accident one in 9,999 times, and consequently he will feel justified in attributing the coincidence to some constant disturbing influence. Similarly, to estimate the evidential value of any other observed uniformity, we only require to know how nearly the odds against chance coincidence have approached to some such standard maximum as 9,999 to 1. But, as any standard must always be more or less arbitrary- some thinking it too lenient and others unnecessarily severe- it is usual to employ a formula giving not the maximum but the middle deviation or "probable error".[1] We may then easily find the probability of mere hazard from the following comparative table:

If the observed correlation divided by the probable error be.	= 1	2	3	4	5	6
then the frequency of occurence by mere hazard	= ½	1/6	1/23	1/143	1/1250	1/19000

Now , the smallness of this probable error depends principally upon the number of cases observed, but also largely upon the mathematical method of correlation. Though a faultiness in the latter respect can theoretically be made good by increasing the range of the observations, yet such increase is not always possible, and, besides,has other grave disadvantages which will be discussed later on. Other things being equal, therefore, <u>the best method is that one which gives the least probable error</u>. For the benefit of the reader, this probable error should always be plainly stated; nothing more is required than a rough approzimation; for while it is highly important to distinguish between a deduction worth, say, 0.9999 of perfect certainty and one worth only 0.75, it would be a mere splitting of straws to care whether a particular experiment works out to a validity of 0.84 or to one of 0.85.

(d) Ease of application.

---

1. In the proper use of this expression.

eliminated, owing to the fact that such accidental deviations are different in every individual case (hence are often called the "variable errors") and occur quite impartially in every direction according to the known laws of probability. The consequence is that they eventually more or less completely counteract one another, and thus finally present an approximately true result. Such elimination, however, must always remain theoretically incomplete, since no amount of such coincidence is absolutely impossible; but beyond certain limits it becomes so extremely unlikely that for practical purposes we can afford to neglect it. When a man loses 11 times running at pitch-and-toss, he can reckon that such a result would not occur by mere accident once in 2,999 times, and consequently he will feel justified in attributing the coincidence to some constant disturbing influence. Similarly, to estimate the evidential value of any other observed uniformity, we only require to know how nearly the odds against chance coincidence have approached to some such standard maximum as 2,999 to 1. But, as any standard must always be more or less arbitrary, some think it too lenient and others unnecessarily severe. It is usual to employ a formula giving not the maximum but the middle deviation or "probable error".[1] We may then easily find the probability of mere hazard from the following comparative table:

If the observed correlation divided by the probable error be.	1	2	3	4	5	6
then the frequency of occurrence by mere hazard	½	1/5	1/22	1/143	1/1350	1/19000

Now, the smallness of this probable error depends principally upon the number of cases observed, but also largely upon the mathematical method of correlation. Such deficiencies in the latter respect can theoretically be made good by increasing the range of the observations, yet such increase is not always possible, and, besides, has other grave disadvantages which will be discussed later on. Other things being equal, therefore, the best method is that one which gives the least probable error. For the benefit of the reader, this probable error should always be plainly stated; nothing more is required than a rough approximation; for while it is highly important to distinguish between a deduction worth, say, 0.9999 of perfect certainty and one worth only 0.75, it would be a mere splitting of straws to care whether a particular expression attains to a validity of 0.84 or to one of 0.85.

(d) Mass of application.

---

1. In the proper use of this expression.

The most accurate ways of calculation are generally somewhat difficult and slow to apply; often, too there occur circumstances under which they cannot be used at all. Hence in addition to a standard method, which must be used for finally establishing the principal results, there is urgent need, also of auxiliary methods capable of being employed under the most varied conditions and with the utmost facility.

But here a word of warning appears not out of place. For such auxiliary methods are very numerous and their results, owing to accidnets, will diverge to some extent from one another; so that the unwary,"self-suggested" experimenter may often be led unconsciously - but none the less unfairly - to pick out the one most favorable for his particular point, and thereby confer upon his work an unequivocality to which it is by no means entitled. Any departure from the recognized standard methods are only legitimate, either when absolutely necessary, or for mere preliminary work, or for indicating comparatively unimportant relations.

2. Standard Methods Explained.

(a) Correlations between variables that can be measured quantitatively.

This may be regarded as the normal type of correlation. Its standard method of calculation is that discovered by Bravais,[1] in 1846, and shown by Pearson in 1896,[2] to be the best possible. Pearson terms this method that of "Product moments."

The formula appears most conveniently expressed as follows:

$$r = \frac{Sxy}{\sqrt{Sx^2 . Sy^2}}$$

where x and y are the deviations of any pair of characteristics from their respective medians,

xy is the product of the above two values for any single individual,

Sxy is the sum of such products for all the individuals,

Sx^2 is the sum of the squares of all the various values of x,

Sy^2 is similarly for y

and r is the required correlation.

A simple example may make this method clearer. Suppose that it was desired to correlate acuteness of sight with that of hearing, and that for this puppose five persons were tested as to the greatest dostance at which they could read and hear a standard alphabet and sound respectively. Suppose the results to be;

---

1. "Memoires par divers savants" T, IX, Paris, pp255-332

2. "Phil. Trans., R.S., London" Vol.CLXXXVII, A, p.164.

The most accurate ways of calculation are generally somewhat difficult and slow to apply; often, too, there occur circumstances under which they cannot be used at all. Hence in addition to a standard method, which must be used for finally establishing the more technical results, there is urgent need, also of auxiliary methods capable of being employed under the most varied conditions and with the utmost facility.

But here a word of warning appears not out of place. For such auxiliary methods are very numerous and their results, owing to [illegible], will diverge to some extent from one another; so that the [illegible], "self-[illegible]" experimenter may often be led incorrectly - all more [illegible] - to [illegible] the one most favorable to his particular point, and thereby confer upon his work an unequivocality to which it is by no means entitled. Any deviations from the [illegible] standard methods are only legitimate, either when absolutely necessary, or for mere preliminary work, or for [illegible] comparatively unimportant [illegible].

2. Standard Methods Explained.

(a) Correlations between variables that are measured quantitatively.

This may be regarded as the normal type of correlation. Its standard method of calculation [illegible] discovered by Bravais in 1846,[1] and shown by Pearson in 1896,[2] to be the best possible. Pearson terms this method that of "product moments."

The formula appears most conveniently expressed as follows:

$$\frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = r$$

where x and y are the deviations of any pair of [illegible] from their respective medians.

xy is the product of the [illegible] two values for any single individual,

Σxy is the sum of such products for all the individuals,

Σx^2 is the sum of the squares of all the [illegible] values of x,

Σy^2 is similarly for y

and r is the required correlation.

A simple example may make this method [illegible]. Suppose that it was [illegible] to correlate [illegible] of sight with that of hearing, and that for this purpose five persons were tested as to the greatest distance at which they could read out a standard alphabet and [illegible] respectively. [illegible] the results to be;

---

1. "Memoires [illegible] divers savants" [illegible], Paris, pp256-332

2. "Phil. Trans., [illegible], London" Vol.CLXXXVII, A, p.164.

Person	Sight	Hearing
A	6ft.	6ft.
B	7	11
C	9 (median)	12
D	11	10 (median)
E	14	8

then, we get

	x	y	xy	x^2	y^2
A	- 3	- 4	+ 12	9	16
B	- 2	+ 1	- 2	4	1
C	0	+ 2	0	0	4
D	+ 2	0	0	4	0
E	+ 5	- 2	- 10	25	4

$$Sxy = \left(\begin{matrix} +12 \\ -12 \end{matrix}\right. \quad Sx^2 = 42 \quad Sy^2 = 25$$

so that $r = \dfrac{+12 - 12}{\sqrt{42 \times 25}} = 0$, and there, thus, is no correspondence, direct or inverse.

The "probable error" between any obtained correlation and the really existing correspondence has been determined by Pearson, as being "with suffieient accuracy"when a fairly large number of cases have been taken,

$$= 0.674506 \sqrt{\frac{1 - r^2}{n\,(1 + r^2)}}$$

For discussion of correlation between characteristics whose distribution differs considerably from the normal probability curve as regards either "range" or "shewness," reference may be made to the works below.[1] It may be remarked that the method of "product moments" is valid, whether or not the distribution follow the normal law of frequency, so long as the "regression" is linear.

(b) Correlation between characteristics that can not be measured quantitavely.

In the example quoted by Galton, of correspondence between the length of arm and that of leg, it may be noted that the correspondence is proportional quantitatively; a long arm has a tendency to be accompanied by a leg not only long, but long to the same degree. Now, in many cases, such proportionality is by the nature of things excluded; a printed work is possibly remembered better than one heard; but, nevertheless, we cannot in accordance with the preceding formula, ascertain whether degrees of visuality are correlated to retentiveness of memory, seeing that in the former case there do not exist any degrees, a word being simply either seen or not seen. Perhaps even

1. Udny Yule: "Proc. R.S. London," Vol. LX, p477
Pearson: "Phil. Trans. R.S. London." Vol.CLXXXV, 1A,p71; Vol CLXXXVI, 1 A, p343, and Vol. CXCI A, p.229.
G. Lipps: "Die Theorie der Collectivgegenstände," Wundt's Phil. Stud., Vol. XVII.

Person	[illegible]	[illegible]
A	6[illegible]	[illegible]
B	[illegible]	11
C	9 (median)	13
D	11	12 (median)
E	14	[illegible]

then, we get

	x	y	xy	x^2	y^2
A	[illegible]	[illegible]	[illegible] 12	9	16
B	[illegible]	[illegible]	[illegible]	4	1
C	0	[illegible]	0	0	4
D	[illegible]	0	0	4	0
E	[illegible]	[illegible]	- 10	25	4

$Sxy = (-12 + 12)$, $Sx^2 = 42$, $Sy^2 = 25$

so that $r = \dfrac{+12 - 12}{\sqrt{42 \times 25}} = 0$, and there, thus, is no correspondence, direct or inverse.

The "probable error" between an obtained correlation and the really existing correspondence has been determined by Pearson, as being, with sufficient accuracy when a fairly large number of cases have been taken,

$$= 0.67[illegible]506 \frac{1 - r^2}{\sqrt{n(1 + r^2)}}$$

For discussion of correlation between characteristics whose distribution differs considerably from the normal probability curve as regards either "range" or "skewness," reference may be made to the works below.[1] It may be remarked that the method of "product moments" is valid, whether or not the distribution follows the normal law of frequency, so long as the "regression" is linear.

(b) Correlation between characteristics that can not be measured quantitatively.

In the example quoted by Galton, of correspondence between the length of arm and that of leg, it may be noted that the correspondence is proportional quantitatively; a long arm has a tendency to be accompanied by a leg not only long, but long to the same degree. But, in many cases, such proportionality is by the nature of things excluded; a printed word is possibly remembered better than one heard; but, nevertheless, we cannot in accordance with the above formula, ascertain whether degrees of visuality are correlated to retentiveness of memory, seeing that in the former case there are no such [illegible] degrees, a word being simply either seen or not seen; perhaps even

---

1. Udny Yule: "Proc. R. S." Vol. LX, p477.
Pearson: "Phil. Trans." [illegible] Vol. CXCV, 1A. p.1; Vol CLXXXVI, [illegible] p.253.
G. Lipps: "Die [illegible] Abhängigkeit," Wundt's Phil. Stud., [illegible]

more numerous are those cases where proportionality does indeed exist, but practically will not admit of being measured; for instance, it is probable that conscientiousness is to some extent a hereditary quality, yet we cannot well directly determine whether brothers tend to possess precisely the same amount of it, owing to the fact that we cannot exactly measure it.

In all such cases we must confine ourselves to counting the frequencies of coexistence. We can easily find out how often seen and spoken words are respectively remembered and forgotten. It has proved quite feasible to divide the children of a school generally into "conscientious" and"non-conscientious," and then to measure how much brothers tend to be in the same division; When we have proved this simple association, we may provisionally assume correlation of quantity also; that is to say, if the "conscientious," generally speaking, have a particular degree of tendency to possess brothers likewise "con scientious," then boys with excessively tender scruples will have the same degree of tendency to possess brothers with similarly excessive tenderness, while those with only a moderate amount of virtue will be thus correlated with brothers also of only moderate virtue; further, the ethical resemblance may be expected to repeat itself in cousins, etc. only reduced in proportion as the kinship is diminished.

For measurement of this non-proportional association, a standard method, which may be termed that of "cross multiples," has been elaborated by Sheppard[1], Bramley-Moore, Filon, Lee, and Pearson. The formula is, unfortunately, too long and complicated to be usefully quoted in this place. It will be found in the under cited work[2] together with its probable error as determined by Pearson.[3] In practice, it will generally have to be replaced by one of the more convenient methods to be next described.

3. Comparison by Rank.

This method of "cross multiples" is not only difficult and tedious of application, but also it gives a probable error nearly double that of "Product moments."

Now, it can often be altogether escaped in the case of quantitiesnnot admitted absolute measurement, by substituting instead <u>comparison</u>. This other way will be discussed at some length, as it has been largely used by myself and is believed chiefly responsible for some successful experiments. All characteristics may be collated from two quite distinct aspects: either (as in example of visual and auditory acuteness) by actual

---

1."Phil. Trans." Vol. CXCII, A, p.141.
2. "Phil. Trans." Vol. CXCV A, pp.2-7
3. Phil. Trans." Vol. CXCV A, 10-14.

mensuzation, or else by order of merit: we might say that a student, A, obtained 8,000 marks in an examination, while B only got 6,000; or instead, we might say that A was third out of 100 candidates, while B was only 20th. Precisely the same method of calculation may be again used in the latter case, simply substituting the inverse ranks, 97, 80, etc., for the performances, 8,000, 6,000, etc.

(a) Disadvantages of the "Rank" method.

In the first place, it may be objected that the observed correlation would then only hold good for persons of the same average differences from one another. For assuming, say, acute sight to be correlated with acute hearing; then the order of merit of A, B and C, as regards sight, is more likely to remain unaltered as regards hearing also, when the difference in their respective powers of vision is extremely marked, than when they are practically equal on the latter head. But the more numerous the persons experimented on, the less will be the average difference of faculty; it might, therefore, be supposed that the correlation would become continually less perfect as the experiments were made more extensive. This, however, would be a fallacy: 100 experimental subjects compared together by "Rank" would on the whole actually show appreciably the same average coffelation as 1,000, provided, that in either case the subjects are selected by chance; the amount of the correlation is not really dependent upon the difference between the grades, but upon the relation of this difference to the mean diviation: and noth of these increase together with the number of subjects; On the other hand, the correlation will undoubtedly diminish if the subjects be all chosen form a more homogeneous class; in a select training school for teachers, fpr example, general intelligence will throughout show smaller correlation with other qualities, than would be the case in a college for quite average young men of the same age; but this fact applies just as much to comparison by "Measurement."

The next possinle objection is that comparison by rank bases itself upon an assumption that all the subjects differ form one another by the same amount, wherwas A amy differ form B five times as much as B differs from C. But such an assumption would only take place, if correspondence by rank were considered to be wholly equivalent to that by measurement; no such assumption is made; the two aspects are recognized to be theoretically distinct. but advantage is taken of the fact that they give correlational values sencibly equivalent in amount. Even against the small existing discrepancy amy be set off a deviation of the same order of magnitude which is incurred when using measurement itself, owing to the practical necessity of throwing the cases into a number of groups.

mensuration, or also by [illegible] student, A, obtained 8,000 marks [illegible] only of 5,000; or indeed, [illegible] of 100 candidates, while B was only 20th. [illegible] method of calculation may be equally used [illegible] simply substituting the [illegible] formances, 8,000, 5,000, etc.

(a) Disadvantages of [illegible].

In the first place, [illegible] relation would then only hold good for [illegible] average differences from the [illegible] eight to be correlated with [illegible] merit of A, B and [illegible] unaltered as regards [illegible] respective powers of [illegible] they are practically equal [illegible] numerous these [illegible] average difference of faculty [illegible] that the correlation [illegible] the experiments were made more extensively [illegible] would be a [illegible] 100 experimental subjects [illegible] gether by [illegible] ciably the same average [illegible] 1,000 [illegible] either case the subjects are selected by [illegible] the correlation is not [illegible] tween the grades, but upon [illegible] the mean division and [illegible] the number of subjects; [illegible] will undoubtedly [illegible] more homogeneous class; in a select training school for teachers, for example, general intelligence will [illegible] correlation with the other qualities, [illegible] college for [illegible] feet [illegible].

The [illegible] objection [illegible] bases itself upon an assumption [illegible] one another [illegible] B five times [illegible] tion would only [illegible] considered to be [illegible] no such assumption is made; [illegible] be theoretically [illegible] they give correlational values [illegible] Even against the small [illegible] deviation [illegible] when [illegible] ity of [illegible]

The third and only solid objection is that rank affords a theoretically somewhat less full criterion of correspondence than does measurement; and the force, even of this argument, disappears on considering that the two methods give apprecibaly the same correlational values.

(b) Advantages of the "Rank" method.

The chief of these is the large reduction of the "accidental error." In normal frequency curve, the outlying exceptional cases are much more spaced apart than are those nearer to the average; hence, any accident disturbing the position of these exceptional cases will have unduly great effect on the general result of the correlation; and owing to this inequality in the influence of the errors, the latter will not compensate one another with the same readiness as usual. Moreover, it is just these hyper-influential extreme cases where there is most likelihood of accidental errors and where there very frequently prevails a law quite different from that governing the great bulk of the cases. As regards the quantity of this gain by using rank (abstracting from the last mentioned point, which cannot well be estimated in any general manner) there should be no difficulty in calculating it mathematically. From a considerable amount of empirical evidence, the probable error when using the method of "product moments" with rank appears to become less than two-thirds of that given by the same method with measurement, and therefore only about one-third of that given by the method of "cross multiples."

The next advantage is that rank eliminates any disparity between the two characteristics compared, as regards their general system of distribution; such a disparity is often not intrinsic or in any way relevant, but merely an effect of the particular manner of gaining the measurement. By means of rank, a series presenting the normal frequency curve can be compared on even terms with another series whose curve is entirely different. This cannot well be done when using measurements. (See p.7).

Rank has also the useful property of allowing any two series to be easily and fairly combined into a third composite one.

(c) Conclusion.

From the practical point of view, it is so urgently desirable to obtain the smallest probable error with a given number of subjects, that the method of rank must often have the preference even when we are dealing with two series of measurements properly comparable with one another.

Theoretically, rank is at any rate preferable to such a hybrid and unmeaning correlation as that between essential measurements on the one side and mere arbitrary classification on the other. As the latter occur in most psychological correlations,

The third and only [illegible]
theoretically somewhat [illegible]
does measurement; and the [illegible] -
ments or considering that the [illegible]
the same correlational [illegible].

(b) Advantages of the "Rank" [illegible].

A brief of these [illegible]
error. In normal [illegible]
cases are much more spaced [illegible]
average; hence, any [illegible]
exceptional cases [illegible]
result of the correlation [illegible]
influence of the error [illegible]
another with the same [illegible]
these hyper-influential [illegible]
liable of accidental [illegible]
prevails a law quite [illegible]
bulk of the cases. In regard [illegible]
[illegible] (abstracting from [illegible]
cannot well be estimated [illegible]
be no difficulty in calculating [illegible]
siderable amount of [illegible]
using the method of [illegible]
to become less than two-thirds [illegible]
with measurement, and [illegible]
given by the [illegible].

The next advantage [illegible]
between the two characteristics [illegible]
general system of [illegible]
trinsic or in any way relevant [illegible]
ticular manner of [illegible]
rank, a [illegible]
compared on [illegible]
entirely different. This [illegible]
measurements. [illegible]

[illegible]
to be [illegible]

(c) [illegible].

[illegible]
to obtain the [illegible]
subjects, [illegible]
once even [illegible]
ments [illegible]

Though [illegible]
and [illegible]
ments [illegible]
others. [illegible]

the only other resource would be to avoid measurements altogether by using the method of "cross multiple." But this trebles the size of the probable error, and therefore renders it necessary that the subjects should be no less than nine, times as numerous; such an enormous increase, even if possible, would generally be accompanied by disadvantages infinitely outweighing the supposed theoretical superiority of method.

The above advantages are still further enhanced whenever dealing with one-sided frequency curves, such as are furnished by most mental tests. For in these cases the great bulk of influence upon the resulting correlation is derived exclusively from the very worst performances and is consequently of a specially doubtful validity.

In short, correlation by rank, in most cases a desirable procedure, is for short series quite indispensable, rendering them of equal evidential value to much longer ones treated by other ways. Luckily, it is precisely in short series that gradation by rank is practically attainable.

(4) Auxiliary Methods.

These, as has been said, are only for use when there is adequate reason for not employing the above "standard" methods. Any number are devisable. Their resulting correlational values do not quite coincide with those found by the standard says, but nearly enough so for most practical purposes.

(a) Auxiliary methods of Pearson

Several very ingenious and convenient ones are furnished by him,[1] but all of similar type and requiring the same data as that of "cross-multiple."[2] They are therefore for use when the compared events do not admit of direct quantitative correlation. The following appears to combine facility and precision to the greatest degree:

$$r = \sin \frac{\pi}{2} \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

where the two compared series of characteristics, say P and Q are each divided into two (preferably about equal) classes; if the case is one where quantity exists but cannot be absolutely measured, P II will comprise the instances in which P is in manifest deficiency; but if the compared characteristics essentially exclude quantity, P II becomes the instances where P is

1. "Phil. Trans. R.S.L.," Vol. CXCV, A, pp 1 and 79.

2. They are all refinements of the original formula, $r = \frac{ad - bc}{ad + bc}$ published by Yule, Proc. R.S.L.," Vol. LXVI, p.23.

the only other resource woul[illegible] be [illegible] gather by using the method of [illegible] trables the size of the probable [illegible] necessary that the subjects sh[illegible] numerous; such or enormous [illegible] generally be accompanied [illegible] ing the supposed theoretical [illegible]

The above advantages [illegible] still further [illegible] dealing with one-[illegible] frequency curve, [illegible] by most mental tests. [illegible] influence upon the [illegible] from the very word [illegible] specially doubtful validity.

In short, correlation by [illegible] cedure, is for short [illegible] of e(m) eviterial value [illegible] were. [illegible] It is [illegible] by [illegible] practically attainable.

(4) Auxiliary Methods.

[illegible] there, as has been said, [illegible] gives reason for not [illegible] above [illegible] Any number are [illegible] do [illegible] quite coincide with [illegible] but nearly enough [illegible]

(5) Auxiliary methods of [illegible]

Several very ingenious [illegible] him,[1] but all of similar [illegible] that of cross-multiple.[2] [illegible] the [illegible] events [illegible] lation. The following [illegible] to the [illegible]

[illegible]

where the two geometrical [illegible] are each divided into [illegible] the curve is [illegible] quadrats [illegible] measured. [illegible] will [illegible] manifest [illegible] ; but [illegible] tially [illegible] quantity. [illegible]

---

1. Phil. Trans. [illegible]

2. They are [illegible] published [illegible]

absent; similarly Q. Then,

a =	the number	of times	that	P I	is accompanied	by	Q I
b "	"	"	"	P II	"	"	Q II
c "	"	"	"	P I	"	"	Q II
d "	"	"	"	P II	"	"	Q II.

If $a + b$ is not very unequal to $c + d$, the probable error may be taken at about $1.1/\sqrt{n}$, where n = the number of instances in the whole of P or of Q.[1]

Returning to our previous illustration, suppose that it was desired positively to ascertain the merits of instruction by writing and by word of mouth respectively. Ten series, each consisting of ten printed words, have been successively shown to a class of twenty children, who each time had to write down by memory as many as they could. The experiment was next repeated, but reading the words aloud instead of showing them. Of the 2,000 visual impressions 900 were corectly remembered, while of the same number of auditory ones only 700 were retained.

Call the visual impressions	P I
" " auditory "	P II
" " remembered "	Q I
" " forgotten "	Q II

then $a = 900$, $b = 700$, $c = 1,100$, $d = 1,300$, and

$$r = \sin \frac{\pi}{2} \; \frac{\sqrt{900} \times \sqrt{1,300} - \sqrt{700} \times \sqrt{1,100}}{\sqrt{900}\;\sqrt{1,300} + \sqrt{700}\;\sqrt{1,100}} = 0.16$$

the probable error then comes to $1.1/\overline{4,000}$ = nearly 0.02, or about 1/8 of the above correlation; so that the latter would not occur by mere chance once in 100,00 times.

We thus see that there is at any rate good prima facie evidence of some superiority on the part of the visual sense. Also, if the experiment has been fairly executed and adequately described, any subsequent verification under sufficiently similar conditions, by other experimenters, should result in a concordant correlation, probably between 0.04 and 0.28.

Moreover, we have obtained a direct estimate of the importance of this apparent superiority of the visual sense; for the square of the correlation amounts to 0.025; so that of the various causes here tending to make the children remember some words better than others, the difference of sense impressed comes to about one fortieth part (see p. 4).

---

1. More accurately, $\sin 0.1686\,\pi\,(1 - r^2)\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

absent; similarly [illegible] Then.

a = the number of times that [illegible] is accompanied by [illegible]
b " " [illegible]
c " " [illegible]
d " " [illegible]

If a + b is not very [illegible] the probable error may be taken as about 1.1/√n, where n = the number of instances in the whole [illegible][1]

[illegible] determining [illegible] previous [illegible] desired positively to [illegible] ing and by word of mouth [illegible] sisting of ten printed [illegible] class of twenty children [illegible] memory as many as they could [illegible] repeated, but reading [illegible] Of the 2,000 visual [illegible] while of the same number [illegible] tained.

Call the visual [illegible]
" " auditory [illegible]
" " remembered [illegible]
" " forgotten [illegible]

then a = 900, b = 700, c = 1,100, d = 1,300, [illegible]

$$r = \sin\frac{\pi}{2}\cdot\frac{\sqrt{900}\times\sqrt{1,300}-\sqrt{700}\times\sqrt{1,100}}{\sqrt{900}\,\sqrt{1,100}\,\sqrt{700}\,\sqrt{1,100}} = [illegible]$$

the probable error then equals 1.1/√[illegible] = [illegible], or about 1/3 of the [illegible] not occur by mere chance [illegible] in 100,000 [illegible].

We thus see that there is [illegible] dence of some [illegible] if the experiment [illegible] scribed, any subsequent verific[illegible] conditions, by other experiments, [illegible] correlation, [illegible] between 0.[illegible] .95.

Moreover, [illegible] tance of [illegible] score of the [illegible] and census [illegible] words both [illegible] comes to [illegible]

---

1. More accurately, [illegible] 1.1 [illegible]

(b) Method of proportional changes.

This is very often convenient, being especially applicable to a large number of psychological experiments, and so easy that the result can be approximately seen on inspection. Here,

$$r = \frac{3}{2} \cdot \frac{a - b}{a - b} \quad {}^{1}$$

where $a =$ the number of cases that have changed in accordance with the supposed correspondence, and $b =$ the number that have changed in contradiction of it. The probable error again comes to $\frac{1.1}{\sqrt{n}}$

Suppose, for example, we were demonstrating that intellectual fatigue may be satisfactorily investigated by the method of Griessbach.[2] With this view, we have applied his test to 100 boys before and after their lessons. In the latter case 68 of them have presented the expected duller sensitivity, but 32, on the contrary, have shown a finer discrimination than before work.

Now, clearly, had the correspondence been perfect, all the hundred would hve become worse.[3] Thus,

$$r = \frac{3}{2} \; \frac{68 - 32}{100} = 0.54.$$

As the probable error comes to 0.11, our imaginary correlation is five times greater, and therefore would not have occurred by mere accident more than once in 1,250 times; so that we become practically certain that the sensitivity of the skin realy does measure fatigue.

It now becomes easy to compare the quantity of this fatigue at different stages of work. Let us say that further experiments, after lessons lasting one hour longer than before, showed the correlation had risen to 0.77. Thereby we see that the influence of fatigue swells from 0.54^2 to 0.77^2, that is, from being 1/5 to being 3/5 of all the sources of variation in cutaneous sensitivity. Such a result has a very different scientific significance from, say, any conclusion that the average sensory threshold had enlarged by so many more millimetres.

---

1. Hence, when the correlation is very complete, say over 0.75 the above formula gives appreciably too large values; as the amount reaches 0.90 and 1, the first factor must be reduced from 3/2 to 5/4 and 1 respectively.

2. This, as is well known, consists in determining the least distance apart at which two points of contact can be distinguished as being double and not single.

3. Assuming, that is to say, that all the boys become fatigued by

Moreover, our test can be easily and precisely compared with any of the various other recommended procedures, being more reliable than all which present smaller correlations and *vice versa*.

(c) Method of class averages.

It often happens that measurements (or ranks) are known but not in such a way as to be able to use either the method of "product moments" or even any of the methods of Pearson. Under such circumstances, I have found it very useful to be able to apply the following relation:

$$r = \frac{d}{D}$$

where d is the observed difference between the average measurement (or rank) of the P's accompanied by Q I and that of those accompanied by Q II, and D is the greatest difference that was possible (such as would have occurred, had the correspondence been perfect). If Q has been divided into two about equal portions, D will be equal to twice the middle or "quartile" deviation from the average in the whole series P; while if Q has been divided after the usual fashion into three such portions, only the two outer ones can be used and then D = 2.87 times the above middle deviation (again taken in the whole series P)

Suppose, for example, that we wish to ascertain whether the well known test of "reaction-time" gives any indication as to the person's general speed of movement. We try a hundred persons both in reaction-time and in speed of running 50 yards. Then we divide the reaction-time records into two classes, I containing all the quickest performers and II al the slowest. We now see how long these two classes of reacters took respectively to run the fifty yards, and what was the middle deviation from the average among all the runners taken together. Let us put the average of class I at 6 seconds, that of class II at 6.5 seconds, and the general middle deviation at 1.1 seconds. Then

$$r = \frac{6.5 - 6}{2 \times 1.1} = 0.23$$

The evidential value of the result is given approximately, even for small values of n, by the following relation:

$$\text{Probable error} = \frac{1.17}{\sqrt{n}} \quad \frac{\sqrt{n} + 1}{\sqrt{n} + 2}$$

where n is the total number of cases considered. In the three-

Moreover, our test can be [illegible] any of the various other recommended procedures, being more reliable than all which give [illegible] correlations and vice versa.

(c) Method of class averages.

It often happens that measurements (or ranks) are known but not in such a way as to [illegible] use either the method of "product moments" or even any of the methods of [illegible].
Under such circumstances, I have found it very useful to be able to apply the following relation:

$$r = \frac{d}{D}$$

where d is the observed difference between the average measurement (or rank) of the P's accompanied by I and that of those accompanied by II, and D is the greatest difference that was possible (such [illegible] would have occurred, had the correspondence been perfect). If [illegible] been divided into two about equal portions, D will be equal to twice the middle or "quartile" deviation from the average in the whole series; while if Q has been divided after the usual fashion into three such portions, only the two outer ones [illegible] used and then D = 2.87 times the above middle deviation (again taken in the whole series.)

Suppose, for example, that we wish to ascertain whether the well known test of "reaction-time" gives any indication as to the person's general speed of movement. We try a hundred persons both in reaction-time and in speed of running 50 yards. Then we divide the reaction-time records into two classes, I containing all the quickest performers and II all the slowest. We now see how long these two classes of persons took respectively to run the fifty yards, and what was the middle deviation from the average among all the runners taken together. Let us put the average of class I at 6 seconds, that of class II at 6.5 seconds, and the general middle deviation at 1.1 seconds. Then

$$r = \frac{6.5 - 6}{2 \times 1.1} = 0.23$$

The evidential value of the result is given approximately, even for small values of n, by the following relation:

$$\text{Probable error} = \frac{1.17}{\sqrt{n}}$$

where n is the total number of cases [illegible] in the three-fold instead of twofold division, the probable error becomes nearly

$$\frac{1.4}{\sqrt{n}} \quad \frac{\sqrt{n}-1}{\sqrt{n}-2}$$

In the above instance, we find that the observed correlation is little over double the probable error; as so much would turn up about once in six times by mere accident, the evidence is not at all conclusive. Therefore we must either observe many more cases - 600 would be necessary to reduce the probable error to 1/5th of the correlation - or else we must find a better method of calculation. If rank had been employed instead of measurement, the evidence would already have been fairly good, and could have been put beyond all reproach by the addition of another 150 observations. If rank had been employed in conjunction with the methof of "product moments" or that of "rank differences," the required smallness of probable error could have been obtained by as few as 36 cases in all!

The method of "class averages" is especially valuable in deciphering the results of other investigators, where the average performances and the middle deviations are usually given (in good work), but not the data required for any of the other methods.

(d) Method of rank differences.

This method appears to deserve mention also, seeing that it seems to unite the facility of the auxiliary methods with a maximum accuracy like that given by "product moments". It depends upon noting how much each individual's rank in the one faculty differs from his rank in the other one; evidently this will be nil when the correlation is perfect, and will increase as the correlation diminishes.[1]

---

1. This general idea seems to have been first due to Binet and Henri ("La fatigue intellectuelle" p.252-261), who, however do not work it out far enough to obtain any definite measure of correlation. Accordingly, Binet makes little further attempt in later research (L'annee psychologique, Vol.IV) to render it of service, and soon appears to have altogether dropped it (L'annee psychologique, Vol. VI.).

   The same isea occurred to myself and was developed as above, without being at the time acquainted with the previous work in this direction by Binet and Henri. In obtaining the above formulae I was greatly assisted by Dr. G. Lipps' showing generally that when an urn contains n balls numbered 1,2,3,..n, respectively; and when they are all drqwn in turn (without being replaced); and when the difference is each time noted between the number on the ball and the order of its drawing; then the most probable (or middle) total sum of such differences, added together without regard to sign, will be

$$= \frac{n^2 - 1}{3}$$

   Previously I had only calculated this value for each particular size of n required by myself. Prof. Hausdorff further showed, generally, that such sum of differences will present a mean square deviation (from the above most probable value)

$$= \sqrt{\frac{(n+1)(2n^2+7)}{45}}$$

fold instead of twofold division, the probable error becomes nearly

$$[illegible]$$

In the above instance, we find that the observed correlation is little over double the probable error; so much would turn up about once in six times by mere accident, the evidence is not at all conclusive. Therefore we must either observe many more cases - 670 would be necessary to reduce the probable error to 1/4th of the correlation - or else we must find a better method of calculation. If ranks had been employed instead of measurement, the evidence would already have been fairly good, and could have been put beyond all doubt by the addition of another 100 observations. If ranks had been employed in conjunction with the method of [illegible] or that of "rank differences," the required smallness of probable error could have been obtained by [illegible] as [illegible]!

The method of 'class averages' [illegible] in deciphering the results of other investigators, where the average performances and the middle deviations are usually given [illegible] good work), but not the data required for any of the other methods.

(c) Method of rank differences.

This method appears to deserve considerable use, seeing that it seems to unite the facility of the auxiliary methods with a maximum accuracy like that given by 'product moments'. It depends upon noting how much each individual's rank in the one faculty differs from his rank in the other one; evidently this will be nil when the correlation is perfect, and will increase as the correlation diminishes.[1]

1. This general idea seems to have been first used by Binet and Henri ('La Fatigue intellectuelle', p.252-261), who, however, do not work it out far enough to obtain any definite measure of correlation. Accordingly, Binet makes little further attempt in later research (L'Année psychologique, vol. IV) to render it of service, and [illegible] dropped it (L'Année psychologique, vol. V).

The same idea occurred to myself and was developed as above, without being at the time acquainted with the previous work in this direction by Binet and Henri. In obtaining the above formulae I was [illegible] Mr. G. [illegible] showing generally that when [illegible] 1,2,3,...,n, respectively; and [illegible] in turn (without being replaced); and when [illegible] is each time noted between the number [illegible] and the order of [illegible]; then the [illegible] total sum of such differences, [illegible] regard to sign, [illegible]

$$\frac{n^2 - 1}{3}$$

Previously I had only calculated this [illegible] for each particular size of n required by myself. Prof. [illegible] further showed, generally, that such sums of differences will present a mean square deviation from the above mean of [illegible]

$$\frac{(n + 1)(2n^2 + [illegible])}{[illegible]}$$

The relation is as follows:

$$R = 1 - \frac{3Sd}{n^2 - 1} \quad [1]$$

where Sd is the sum of the differences of rank for all the individuals,
n is the total number of individuals,
and R is the required correlation.

The probable error will then be approximately, even for small values of n, $= 0.4/\sqrt{n}$.

To take again the example from p.9, we number the five persons according to their order of merit in hearing and seeing respectively.

person	seeing rank	hearing rank	difference
A	1	1	0
B	2	4	2
C	3	5	2
D	4	3	1
E	5	2	3
			Sd = 8

so that

$$R = 1 - \frac{3 \quad 8}{25 - 1} =$$

and again we find that there is no correlation, direct or inverse.

This method, though very accurate and pre-eminently quick in application, has unfortunately four serious disadvantages.

It can be only used for ranks, and not immediately for measurements.

The probable error given is only that showing how great correlations may be expected from pure accident when there is no really existing correspondence between the two characteristics. It does not (like Pearson's probable error for the method of "product moments") directly show how much the observed correlation may be expected to differ by accident from any correspondence that does exist.

The various possible values of Sd are found to fall into a frequency curve of marked asymmetry; so that we cannot (as in all the other methods here given) take the minus values of R as representing so much inverse correlation. This defect could be remedied mathematically; but there are also other respects in which this side of the frequency curve appears unsuitable for our purpose, so that it is better to treat every cor-

---

1. This formula becomes slightly incorrect, whenever two or more individuals are bracketed as having precisely the same rank; but the consequent error is usually to be worth considering.

The relation is as follows:

$$R = 1 - \frac{3\Sigma d}{n^2 - 1}$$

where Σd is the sum of the differences of rank for all the individuals,
n is the total number of individuals,
and R is the required correlation.[1]

The probable error will then be approximately, even for small values of n, $= 0.43/\sqrt{n}$.

To take again the example from p.9, we number the five persons according to their order of merit in hearing and seeing respectively.

person	seeing rank	hearing rank	difference
A	1	1	0
B	2	4	2
C	3	5	2
D	4	3	1
E	5	2	3

$\Sigma d = 8$

so that

$$R = 1 - \frac{3 \times 8}{25 - 1} = 0$$

and again we find that there is no correlation, direct or inverse.

This method, though very accurate and pre-eminently quick in application, has unfortunately four serious disadvantages.

It can be only used for ranks, and not immediately for measurements.

The probable error given is only that showing how great correlations may be expected from pure accident when there is no really existing correspondence between the two characteristics. It does not (like Pearson's probable error for the method of "product moments") directly show how much the observed correlation may be expected to differ by accident from any correspondence that does exist.

The various possible values of R are found to fall into a frequency curve of marked asymmetry; so that we cannot (as in all the other methods here given) take the minus values of R as representing an exact inverse correlation. This defect could be remedied mathematically; but there are also other respects in which this side of the frequency curve appears unsuitable for our purpose, so that it is better to treat every cor-

---

1. This formula becomes slightly incorrect, whenever two or more individuals are bracketed as having precisely the same rank; but the consequent error is usually too small to be worth considering.

relation as positive (which can always be done by, if necessary, invertèng the order of one of the series).

Finally, this value R is not numerically equivalent to the "r" found by all the other methods, but for chance distributions appears $\pm\sqrt{r^{3}}$. So far, the proof of this relation is only empirical, but it resśs on a large number of cases taken, however, only between o.20 and 0.60. If it be accepted r can at once be found from the following table;

R	0.05	0.10	0.20	o.30	0.40	0.50	0.60	0.70	0.80	0.90	1
r	o.13	0.22	0.34	0.44	0.54	0.63	0.71	0.79	0.86	0.93	1

Part II.
Correction of "Systematic Deviations."

1.Systematic Deviations Generally.

In the first part, we have seen that any correlational experiments however extensive, can only be regarded as a"sample" out of the immense reality, and will consequently present a certain amount of accidental deviation from the real general tendency; we have further seen that this accidental deviation is measurable by the "probable error" whose determination, therefore becomes an indispensab;e requisite to all serious research.

But now we are in danger of falling from Scylla into Charybdis. For after laboriously compiling sufficient cases and conscientiously determining the probale error, there exists a very human tendency to cease from labor and inwardly rejoice at having thus risen from common fallacious argument tothe serene certainty of mathematics. But whether or not such complacency may be justifiable in pure statistical inquiry, it is at any rate altogether premature in the kind of research that we are at present contemplating; we are not dealing with statistics, but with a line of work so fundamentally different that it may be aptly distinguished by the term of "statisticoids." Here the accidental deviation is not the sole one, nor even the most momentous; there are many other enemies who are unmoved by the most formidable array of figures. These consist in such deviations as, instead of merely being balanced imperperfectly, lie wholly pm the one side or the other. As in ordinary measurements, so too in correlation, we may speak, not only of "accidental" " variable" or "compensating" inaccuracies, but also of "systematic,2 "constant," or "non-compensating" ones.

These systematic deviations are of very varied nature, the most insidious being as usual self suggestion. To take, for instance, one of our recent examples, suppose that we have applied the Griessbach test to a number of children before and

after their lessons, and have found the desired correlation between fatigue and cutanuous insensitivity it still remains exceedingly difficult to convince ourselves that we executed our tests entirely without favor or affection; for it is almost impossible to determine a series of sensory thresholds without some general tendency, either to bring them towards the desired shape, or else - endeavoring to escape such bias- to force them in the opposite direction. To convince others of our impartiality may be harder still. Even this sort of deviation is to be remedied by our proposed exact method of procedure for by it we obtain perfectly definite results which any impartial experimenters may positively corroborate or refute.

2. "Attenuation" by Errors.

From page 3 it will be obvious that a correlation does not simply depend on the amount of concording factors in the two compared series, but solely on the proportion between those concording elements on the one hand and the discording ones on the other. In our examplem it did not matter whether A and B each had one pound or a thousand pounds in the common funds; but only whether the amount was a small or large fraction of their whole incomes. If the discordance, 1-x, be nil, then the concordance x is thereby perfect, that is, = 1; and if the influence of the discordant elements be sufficiently increased, then any concordance will eventually become infinitely small.

To consider a still more concrete example suppose three balls to be rolled along a well-kept lawn; then the various distances they go will be almost perfectly correlated to the various forces with which they were impelled. But let these balls be cast with the same inequalities of force down a rough mountain side; then the respective distances eventually attained will have[1] but faint correspondence to the respective original momenta.

Thus it will be clear that here the accidental deviations have a new consequence simultaneous with, but quite distinct from, that discussed in the last chapter. For there, they impartially augmented and diminished the correlation, tending in a prolonged series to always more and more perfectly counterbalance one another; and in ordinary measurements, this is their sole result. But here in correlations, they also have this new effect which is always in the direction of "attenuating" the apparent correspondence and whose amount, depending solely on the size of the middle error, cannot be in the least eliminated by

---

I. This fact has already been mathematically expressed in the last chapter by the value of correlation between two series being proportional (inversely) to the value of the middle deviation inside the series (see p. 15).

after their lessons, and have found the desired correlation between fatigue and cutaneous insensitivity it still remains exceedingly difficult to convince ourselves that we executed our tests entirely without favor or affection; for it is almost impossible to determine a series of sensory thresholds without some general tendency, either to bring them towards the desired shape, or else - endeavoring to escape such bias - to force them in the opposite direction. To convince others of our impartiality may be harder still. Even this sort of deviation is to be remedied by our proposed exact method of procedure for by it we obtain perfectly definite results which any impartial experimenters may positively corroborate or refute.

2. "Attenuation" by Errors.

From page 7 it will be obvious that a correlation does not simply depend on the amount of concording factor in the two compared series, but solely on the proportion between the concording elements on the one hand and the discording ones on the other. In our example it did not matter whether A and B each had one pound or a thousand pounds in the common funds but only whether the amount was a small or large fraction of their whole incomes. If the discordance, 1-x, be nil, then the concordance x is perfectly perfect, that is, = 1; and if the influence of the discordant elements be sufficiently increased, then any concordance will eventually become infinitely small.

To consider a still more concrete example, suppose three balls to be rolled along a well-kept lawn; then the various distances they go will be almost perfectly correlated to the various forces with which they were impelled. But let these balls be cast with the same inequalities of force down a rough mountain side; then the respective distances eventually attained will have but faint correspondence to the respective original moments.[1]

Thus it will be clear that here the accidental deviations have a new consequence simultaneous with, but quite distinct from that discussed in the last chapter. For there, they impartially augmented and diminished the correlation, tending in a prolonged series to always more and more perfectly counterbalance one another; and in ordinary measurements, this is their sole result. But here in correlations, they also have this new effect which is always in the direction of "attenuating" the apparent correspondence and whose amount, depending solely on the size of the middle error, cannot be in the least eliminated by

1. This fact has already been mathematically expressed in the last chapter by the value of correlation between two series being proportional (inversely) to the value of the middle deviation inside the series (see p. 15).

any prolongation of the series. The deviation has thus become general or "systematic."

Now suposse that we wish to ascertain the correspondence between a series of values, p, and another series, q. By practival observation we evidently do not obtain the true objective values, p and q, but only approximations which we will call p' and q'. Obviously, p' is less closely connected with q' than is p with q. for the first pair only correspond at all by the intermediation of the second pair; the real correspondence between p and q shortly r_{pq}, has been "attenuated" into $r_{p'q'}$.

To ascertain the amount of this attenuation and thereby discover the true correlation it appears necessary to make two or more independent series of observations of both p and q. Then

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}$$

where $r_{p'q'}$ = the mean of the correlations between each series of values obtained for p with each series obtained for q.

$r_{p'p'}$ = the average correlation between and another of these several independently obtained series of values for p.

$r_{q'q'}$ = the same as regards q.

and r_{pq} = the required real correlation between the true objective values of p and q.

Thus, if for each characteristic two such independent series of observations be made, say $p_1 p_2 q_1$ and q_2 then the true

$$r_{pq} = \frac{r_{p1q1} + r_{p1q2} + r_{p2q1} + r_{p2q2}}{4\sqrt{(r_{p1p2} \times r_{q1q2})}}$$

Should circumstances happen to render say p1 much more accurate than p2 then the correlations involving p1 will be considerably greater than those involving p2. In such case, the numerator of the above fraction must be formed by the geometrical instead of by the arithmetical mean; here by the accidental errors of the respective observations cease to eliminate one another and therefore double their final influence; they also introduce an undue diminuaion of the fraction.[1]

In some exceptional and principally very theoretical cases, it may happen that either of the actual measurements, say p'1 is

---

1 By an inversion of the above formula, the correlation between two series of oberavations will be found a useful measure of the accuracy of the observation.

any prolongation of the series. The deviation thus become general or "systematic."

Now suppose that we wish to ascertain the correspondence between a series of values, p, and another series, q. By practical observation we evidently do not obtain the true objective values, p and q, but only approximations which we will call p' and q'. Obviously, p' is less closely connected with q' than is p with q, for the first pair only correspond through the intermediation of the second pair; the real correspondence between p and q, shortly r_{pq}, has been "attenuated" into $r_{p'q'}$.

To ascertain the amount of this attenuation, and thereby discover the true correlation, it appears necessary to make two or more independent series of observations of both p and q. Then

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}$$

where $r_{p'q'}$ = the mean of the correlations between each series of values obtained for p with each series obtained for q.

$r_{p'p'}$ = the average correlation between one and another of these several independently obtained series of values for p.

$r_{q'q'}$ = the same as regards q.

and r_{pq} = the required real correlation between the true objective values of p and q.

Thus, if for each character the two such independent series of observations be made, say p_1, p_2 and q_1, q_2, then the true

$$r_{pq} = \frac{r_{p_1q_1} + r_{p_1q_2} + r_{p_2q_1} + r_{p_2q_2}}{4\sqrt{(r_{p_1p_2} \cdot r_{q_1q_2})}}$$

Should circumstances happen to render [illegible] much more accurate than [illegible], then the correlations involving [illegible] will be considerably greater than those involving [illegible]. In such case, the numerator of the above formula must be formed by the geometrical instead of by the arithmetical mean; [illegible] accidental errors of the respective observations tend to eliminate one another and therefore double their real influence, [illegible] also introduce an undue diminution of the fraction.[1]

In some [illegible] and principally in [illegible] it may happen that [illegible] the actual [illegible]

1 By an inversion of the above formula, the correlation between two series of observations will be found a useful measure of the accuracy of the observation.

connected with q'(or q) quite independently of p or any other link common to p'2. Then, the correlation $r_{p'q'}$ will be to that extent increased without any proportional increase in $r_{p'p'}$; hence our above formula will fallaciously present too large a value.

A greater practical diffuculty is that of obtaining two series sufficiently independent of one another. For many errors are likely to repeat themselves; even two separate observers are generally, to some extent warped by the same influences; we are all imposed on by, not only the "Idola Specus," but alsothe "Idola Tribus" and the "Idola Fori." In such case, the above formula is still valid, only, its correction does not go quite far enough, - a fallacy at any rate on the right side.

An actual instance will best show the urgent necessity of correcting this attenuation. In a correlation between two events, say P and Q, I obtained three independent observations both of P and of Q. The average correlation for those of P with those for Q was o.38 ($= r_{p'q'}$); the average correlation of those for P with one another was 0.58 ($= r_{p'p'}$), the same for Q was 0.22 ($= r_{q'q'}$). Therefore, the correspondence between the real events, P and Q, comes by reckoning to

$$\frac{0.38}{\sqrt{0.58 \times 0.22}} = \text{approximately } 1;$$

so that the correspondence instead of being merely 0.38, appeared to be absolute and complete.

Attenuation by errors can alsobe corrected in another manner, which has the great advantage of an independent empirical basis, and therefore of not being subject to either of the two above mentioned fallacies besetting the other method. Hence, when the results coincide both ways, the fallacies in question may thereby be considered as disproved, for it is very unlikely that they should noth be present and in such proportions as to exactly cancel one another. In this method, instead of directly employing the values $p_1 p_2 p_3$, etc, we amalgamate them into a single list; by this means we clearly eliminate some portion of the individual observational errors, and thereby we cause any really existing correspondence to reveal itself in greater completeness. Now, this increase in correlation from this partial elimination of errors will furnish a measure of the increase to be expected from an entire elimination of errors. Assuming the mean error to be inversely proportional both to this increase in the correlation and to the square root of the number of lists amalgamated, the relation will be:

$$r_{pq} = \frac{\sqrt[4]{mn}\cdot r_{p''q''} - r_{p'q'}}{\sqrt[4]{mn} - 1}$$

connected with q'(or q) quite independently of p or any other link common to p'Q. Then, the correlation $r_{p'q'}$ will be to that extent increased without any proportional increase in $r_{p'p''}$; hence our above formula will fallaciously present too large a value.

A greater practical difficulty is that of obtaining two series sufficiently independent of one another. For many errors are likely to repeat themselves; even two separate observers are generally, to some extent warped by the same influences; we are all imposed on by, not only the "Idola Specus," but alsothe "Idola Tribus" and the "Idola Fori." In such case, the above formula is still valid, only, its correction does not go quite far enough, - a fallacy at any rate on the right side.

An actual instance will best show the urgent necessity of correcting this attenuation. In a correlation between two events, say P and Q, I obtained three independent observations both of P and of Q. The average correlation for those of P with those for Q was 0.38 ($= r_{p'q'}$); the average correlation of those for P with one another was 0.58 ($= r_{p'p''}$); the same for Q was 0.32 ($= r_{q'q''}$). Therefore, the correspondence between the real events, P and Q, comes by reckoning to

$$\frac{0.38}{\sqrt{0.58 \times 0.32}}$$

approximately 1; so that the correspondence instead of being merely 0.38, appeared to be absolute and complete.

Attenuation by errors can also be corrected in another manner, which has the great advantage of an independent empirical basis, and therefore of not being subject to either of the two above mentioned fallacies besetting the other method. Hence, when the results coincide both ways, the fallacies in question may thereby be considered as disproved, for it is very unlikely that they should both be present and in such proportions as to exactly cancel one another. In this method, instead of directly employing the values $P_1 P_2$, etc, we amalgamate them into a single list; by this means we accidentally eliminate some portion of the individual observational errors, and thereby we cause any really existing correspondence to reveal itself in greater completeness. Now, this increase in correlation from this partial elimination of errors will furnish a measure of the increase to be expected from an entire elimination of errors. Assuming the mean error to be inversely proportional both to this increase in the correlation and to the square root of the number of lists amalgamated, the relation will be:

$$\frac{\sqrt[4]{m.n}\cdot r_{p''q''} - r_{p'q'}}{\sqrt[4]{m.n} - 1} = r_{pq}$$

connected with q' (or q) quite independently of p or any other link common to p. Then, the correlation $r_{p'q'}$ will be to that extent increased without any proportional increase in $r_{p'p'}$; hence our above formula will fallaciously present too large a value.

A greater practical difficulty is that of obtaining two series sufficiently independent of one another. For many errors are likely to repeat themselves; even two separate observers are generally, to some extent, warped by the same influences; we are all imposed on by, not only the "Idola Specus," but also the "Idola Tribus" andthe "Idola Fori." In such cases, the abobe formula is still valid, only its correction does not go quite far enough,- a fallacy at any rate on the right side.

An actual instance will best show the urgent necessity of correcting this attenuation. In a correlation between two events, say P and Q, I obtained three independent observations both of P and of Q. The average correlation for those of P with those for Q was 0.38 ($= r_{p'q'}$); the average correlation of those for P with one another was 0.58 ($= r_{p'p'}$); the same for Q was 0.22 ($= r_{q'q'}$). Therefore, the correspondence between the real events, P and Q comes by reckoning to $\frac{0.38}{\sqrt{0.58 \times 0.22}}$ = approximately 1; so the the correspondence instead of being merely 0.38, appeared to be absolute and complete.

Attenuation by errors can also be corrected in another manner, which has the great advantage of an independent empirical basis, and therefore of not being subject to either of the two above mentioned fallacies besetting the other method. Hence, when the results coincide both ways, the fallacies in question may thereby be considered as disproved, for it is very unlikely that they should both be present and in such proportions as to exactly cancel one another. In this method, instead of directly employing the values p_1 p_2 p_3, etc. we amalgamate them into a single list; by this means we clearly eliminate <u>some portion</u> of the individual observational errors, and thereby we cause any really existing correspondence to reveal itself in greater completeness. Now, this increase in correlation from this partial elimination or errors will furnish a measure of the increase to be expected from an entire elimination of errors. Assuming the mean error to be inversely proportional both to this increase in the correlation and to the square root of the number of lists amalgamated, the relation will be:

$$r_{pq} = \frac{\sqrt[4]{mn.r_{p''q''}} - r_{p'q'}}{\sqrt[4]{mn} - 1}$$

connected with q_1 (or q) quite independently of p or any other link common to p_1. Then, the correlation $r_{p'q'}$ will be to that extent increased without any proportional increase in $r_{p'p'}$; hence our above formula will fallaciously present too large a value.

A greater practical difficulty is that of obtaining two series sufficiently independent of one another. For many errors are likely to repeat themselves; even two separate observers are generally, to some extent, warped by the same influences; we are all imposed on by, not only the "Idola Specus," but also the "Idola Tribus" and the "Idola Fori." In such cases, the above formula is still valid, only its correction does not go quite far enough, - a fallacy at any rate on the right side.

An actual instance will best show the urgent necessity of correcting this attenuation. In a correlation between two events, say P and Q, I obtained three independent observations both of P and of Q. The average correlation for those of P with those for Q was 0.38 ($= r_{p'q'}$); the average correlation of those for P with one another was 0.58 ($= r_{p'p'}$); the same for Q was 0.22 ($= r_{q'q'}$). Therefore, the correspondence between the real events, P and Q comes by reckoning to

$$\frac{0.38}{\sqrt{0.58 \times 0.22}} = \text{approximately } 1;$$

so the the correspondence instead of being merely 0.38, appeared to be absolute and complete.

Attenuation by errors can also be corrected in another manner, which has the great advantage of an independent empirical basis, and therefore of not being subject to either of the two above mentioned fallacies besetting the other method. Hence, when the results coincide both ways, the fallacies in question may thereby be considered as disproved, for it is very unlikely that they should both be present and in such proportions as to exactly cancel one another. In this method, instead of directly employing the values $p_1\ p_2\ p_3$, etc. we amalgamate them into a single list; by this means we clearly eliminate <u>some</u> portion of the individual observational errors, and thereby we cause any really existing correspondence to reveal itself in greater completeness. Now, this increase in correlation from this partial elimination of errors will furnish a measure of the increase to be expected from an entire elimination of errors. Assuming the mean error to be inversely proportional both to this increase in the correlation and to the square root of the number of lists amalgamated, the relation will be:

$$r_{pq} = \frac{\sqrt[4]{mn}\ r_{p''q''} - r_{p'q'}}{\sqrt[4]{mn} - 1}$$

where m and n = the number of independent gradings for p and q respectively,

$r_{p'q'}$ = the mean correlation between the various gradings for p and those for q,

and $r_{p''q''}$ = the correlation of the amalgamated series for p with the amalgamated series for q.

In the above quoted instance, the three observations for series P were amalgamated into a single list, and similarly those for series Q. Upon this being done, the two amalgamated lists now presented a correlation with one another of no less than 0.66(= $r_{p''q''}$). Thus by this mode of reckoning, the real correspondence became

$$= \frac{\sqrt[4]{3 \times 3 \times 0.66} - 0.38}{\sqrt[4]{3 \times 3} - 1} = \text{once more approximately } 1,$$

so that this way also the correspondence advanced from 0.38 to absolute completeness.

If more than two independent series of observations are available, we may acquire additional evidence by trying the effect of partial amalgamtion. Instead of throwing all our obtained values together, we may form a set of smaller combinations for each of the two compared characteristics, and then see the mean correlation between one set and the other. In our obove instance instead of summarily considering $p'_1\ p'_2\ p'_3$, we can have $p'_1\ p'_2\ p'_1\ p'_3$ and $p'_2\ p'_3$ and find out their mean correlation with similar values for q. This works out actually to 0.55. Hence

$$r_{pq} = \frac{\sqrt[4]{2 \times 2 \times 0.55} - 0.38}{\sqrt[4]{2 \times 2} - 1} = \text{approximately } 1.$$

Thus again , by this third may, where both terms are the mean of 9 observed correlational values, the correspondence once more rises from the apparent 0.38 to the real 1. (1)

3. Limits of Associative Problems.

We have seen that "the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long leg and conversely;" also that this correlation is defined mathematically by any constant which deter-

---

(1) The exactness of the coincidence between the two methods of correction is in the above instance neither greater nor less than generlly occurs in practice. It was singled out, in order to show that the formulae still hold perfectly good even for such an enormous rise as from 0.38 to 1. The posibility of such a rise is due to the unusual conditions of the experiment in question, whereby the three observations of the same objective series presented the extraordinarily small intercorrelation of 0.22.

where m and n = the number of independent gradings for p and q respectively.

$r_{p'q'}$ = the mean correlation between the various gradings for p and those for q,

and $r_{p''q''}$ = the correlation of the amalgamated series for p with the amalgamated series for q.

In the above quoted instance, the three observations for series P were amalgamated into a single list, and similarly those for series Q. Upon this being done, the two amalgamated lists now presented a correlation with one another of no less than 0.66($= r_{p''q''}$). Thus by this mode of reckoning, the real correspondence became

$$= \frac{\sqrt[4]{3 \times 3} \times 0.66 - 0.38}{\sqrt[4]{3 \times 3} - 1} = \text{once more approximately } 1,$$

so that this way also the correspondence ady need from 0.38 to absolute completeness.

If more than two independent series of observations are available, we may acquire additional evidence by trying the effect of partial amalgamation. Instead of throwing all our obtained values together, we may form a set of smaller combinations for each of the two compared characteristics, and then see the mean correlation between one set and the other. In our above instance instead of summarily considering $p'_1\, p'_2\, p'_3$, we can have $p'_1 p'_2$, $p'_1 p'_3$ and $p'_2 p'_3$ and find out their mean correlation with similar values for q. This works out actually to 0.55. Hence

$$r_{pq} = \frac{\sqrt[4]{2 \times 2} \times 0.55 - 0.38}{\sqrt[4]{2 \times 2} - 1} = \text{approximately } 1.$$

Thus again, by this third way, where both terms are the mean of 9 observed correlational values, the correspondence once more rises from the apparent 0.38 to the real 1. (1)

3. Limits of Associative Problems.

We have seen that "the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long leg and conversely;" also that this correlation is defined mathematically by any constant which deter-

---

(1) The exactness of the coincidence between the two methods of correction is in the above instance neither greater nor less than generally occurs in practice. It was singled out, in order to show that the formulae still hold perfectly good even for such an enormous rise as from 0.38 to 1. The possibility of such a rise is due to the unusual conditions of the experiment in question, whereby the three observations of the same objective series presented the extraordinarily small intercorrelation of 0.22.

mines the function of any definite size of arm to the mean of the sizes of the corresponding legs. These terms, taken literally, are very wide reaching and express what we will call the "universal" correlation between the two organs.

But evidently not the most painstaking investigation can possibly secure any adequately regrresentative sample for such universal correlations, even in the simple case of arms and legs. To begin with, they would have to be equally derived from every stage of growth, including the prenatal period; since this is the most influencial of all canses of variation in size. In the next place, they would have to come from every historical epoch, containing their fair proportion of big Cro-Magnons, little Futfoozers, etc. Further, they must impartially include every living race, from the great Patagonians to the diminutive M'Kabbas; also every social class, from the tall aristocrats to the undersized slummers.

Practically, then, the universal correlation, even if desirable, is quite inaccessible. We are forced to successively introduce a large number of restrictions: the sample is confined to adults, to moderns, to some particular country, etc., etc. In a word, we are obliged to deal with a special correlation.

When we proceed to more narrowly consider these restrictions, it soon becomes clear that they are far from being really detrimental. For every serious investigation will be found to be directed, however vaguely and unconsciously, by some hypothesis as to the causes both of the correspondence and of the digression therefrom. (see page 74). This hypothesis will determine a particular system of restriction, such as to set the correspondence in the most significant relief.

But from these restrictions will at the same time proceed several kinds of grave errors. In the first place, since the restrictions are not explicitly recognized, they often are not carried out in a manner scientifically profitable; then, the result, however true, may nevertheless be trivial and unsuggestive. For instance, a series of experiments was recently executed by one of our best known psychologists and ended - to his apparent satisfaction - in showing that some children's school-order was largely, correlated with their height, weight and strength. As, however, no step had been taken to exclude the variations due to difference of age, the only reasonable conclusion seemed to be that as children grow older they turned out in fact to probably be the true and sufficient one.

The next fault to be feared is equivocality. For even if the controlling underthought be good, yet its indistinctness in the mind of the experimenter caused the restriction to be carried out so unsystematically, that the results inevitable become ambiguous and fruitless.

The last is that, even with the clearest purpose, this specialization of the correlation is an exceedingly difficult matter to execute successfully/ Only by a profound knowledge of the many factors involved, can we at all adequately exclude those irrelevant to out main intention.

Now, all such elements in a correlation as are foreigh to the investigator's explicit or implicit purpose will, like the attuating errors, constitute imputities in it and will quantitatively falsify its appraent amount. This will chiefly happen in two ways.

Minus the function of [illegible]
corresponding law. These [illegible]
express what [illegible]

[illegible] can
possibly [illegible]
even [illegible] to be
equally [illegible]; since
this [illegible] the next
place, [illegible] their
properties [illegible] must [illegible]
tally include every [illegible]
M'Fabben also [illegible]
[illegible]

[illegible]
in other [illegible]
of [illegible]
country, etc., [illegible]

When [illegible]
clear [illegible]
gation [illegible]
hypot[illegible]
sion. (see page 76). [illegible]
restriction, [illegible]

[illegible]
several [illegible]
explicit [illegible]
profitable; [illegible]
unscientific [illegible]
of our best [illegible]
that some [illegible]
and strange [illegible]
to different [illegible]
grow [illegible]

The next [illegible]
thought [illegible]
restriction [illegible]
mass [illegible]

[illegible]
specialization [illegible]
successfully [illegible]
of all [illegible]

[illegible] foreign to the
[illegible]
constitute [illegible]
This will [illegible]

4. "Constriction"and"Dilation."

Any correlation of either of the considered characteristics will have been admitted irrelevantly, if it has supervened irrespectively of the original definition of the correspondence to be investigated. The variations are thereby illegitimately constrained to follow some irrelevant direction so that(as in the case of Attenuation) they no longer possess full amplityde of possible correlation in the investigated direction; the maximum instead of being 1 will be only a fraction, and all the lesser degrees of correspondence will be similarly affected; such a falsification may be called "constriction." Much more rarely, the converse or "dilation" will occur, by correlations being irrelevantly excluded. The disturbance is measureable by the following relation:

$$r_{pq} = \frac{r_{pq}}{\sqrt{1 - r^2_{pv.}}}$$

where r_{pq} = the apparent correlation of p and q. the two variables to be compared,

r_{pv} = the correlation of one of the above variables with a third and irrelevantly admitted variable v.

and r_{pq} = the real correlation between p and q after compensating for the ligitimate influence of v.

Should any further irrelevant correlation, say r_{pw}, be admitted, then

$$r_{pq} = \frac{r^1_{pq}}{\sqrt{1 - r^2_{pv.} - r^2_{pw}}}$$

In the reverse case of "dilation,"

$$r_{pq} = r^1_{pq} \sqrt{1 - r^2_{pv} - r^2_{pw}}$$

these formulae will be easily senn to be at once derivable from the relations stated on pages 74 and 75. Small, irrelevant variations evidently do not affect the result in any sensible degree, while large ones are capable of revolutionizing it.

The following is an actual illustration of this constriction. I was investigating the correspondence between on the one hand intelligence of school lessons and on the other the faculty of discrimination of musical pitch. The correlation proved to be 0.49. But, upon inquiry, it turned out that more than half of the children took lessons in music and there fore enjoyed artificial training as regards pitch; here, then, was a powerful cause of variation additional and quite irrelevant to the research, which dealt with the correspondence between the two natural faculties. When this disturvact had once been ditected, there was no difficulty in eliminating its influence by the above formula; the correspondence between pitch ~~and~~ discrimination and music lessons was measured at 0.61; so that the true required correlation became

$$\frac{0.49}{\sqrt{1 - 0.61^2.}} + 0.62.$$

In this particular case, the more desirable course was open of elimination the constriction, practically, by confining the experiment to those children who were learning music and therefore were on a sufficient equality as regards the training The correlation then gained in this purely empirical way exactly coincided with the former results, being again 0.62.

5. "Distortion."

Whereas Attention and Constriction have wholly tended to reduce the apparent correlation, and Dilation to enlarge it, we now come to a third kind of impurity that may equally well reduce or enlarge. Its effects is thus analogous to the first consequence of accidental errors discussed in the first part of this article, but unlike the latter, this Distortion does not in the least tend to eliminate itself in the longest series of observations.

Distortion occurs whenever the two series ~~of~~ to be compared together both correspond to any appreciable degree with the same third irrevant variant. In this case, the relation is given by

$$r_{pq} = \frac{r^1_{pq} - r_{pv} \cdot r_{qv}}{\sqrt{(1 - r^2_{pv})(1 - r^2_{qv})}} \qquad (1)$$

where $r^1\,pq$ = the apparent correlation between p and q, the two characteristics to be compared,

r_{pv} and r_{qv} = the correlations of p and q with some third and perturbing variable v.

and r_{pq} = the required real correlation between p and q after compensating for the illigitimate influence of v.

Should the common correspondence with v have been irrelevantly excluded instead of admitted, the relation becomes

$$r_{pq} = r^1\,pq \cdot \sqrt{(1 - r^2\,pv)(1 - r^2\,qv)} + r_{pv} \cdot r_{qv}$$

In the course of the same investigation above alluded to, but in another school, the correlation between school intelligence and discrimination of pitch turned out to be - 0.25, so that apparently not the cleverer but the stupider children could discriminate best! But now it was observed that a superiority in discrimination had been shown by the older children, amounting to a correlation of 0.55; while, for a then unknown reason, the schoolmaster's estimate of intelligence had shoen a very marked (though unconscious) partiality for the younger ones, amounting to a correlation of 0.65. Hence the true correlation reckoned out to

$$\frac{-0.25 - 0.55 \times (-0.65)}{\sqrt{(1 - 0.55^2)(1 - -0.65^2)}}$$

= + 0.17. This latter low but direct correlation was - under the particular

1. This same formular has already been arrived at, ~~through~~ though along a very different route, by Yule. See Proc. R.S.L. Vol. LX.

In this particular case, [illegible] constellation, [illegible] learning [illegible] the [illegible] former results, [illegible].

5. "Distortion."

[illegible] apparent [illegible] of inquiry [illegible] the first consequence of [illegible] article, [illegible] itself in the [illegible].

Distortion occurs [illegible] both correspond to any [illegible] same [illegible]. In this case, [illegible] is given by

$$r_{[illegible]} = \frac{[illegible]}{\sqrt{1 - [illegible]\,(1 - [illegible])}} \qquad (1)$$

where [illegible] to [illegible]

r_{xy} and [illegible] pertaining [illegible]

and $r_{[illegible]}$ = the [illegible] for the [illegible]

Should [illegible] excluded instead of [illegible]

$$[illegible]$$

In [illegible] in another school, [illegible] of pitch turned out [illegible] significance in discrimination had been shown by [illegible] of 0.[illegible]; while, for [illegible] a very marked [illegible] amounting to a correlation of 0.[illegible]. [illegible]

$$\frac{[illegible] - 0.[illegible]}{\sqrt{(1 - [illegible])\,(1 - 0.[illegible])}}$$

= + 0.[illegible]. [illegible] particular

1. This [illegible], through [illegible] a very different [illegible], Vol. IX.

circumstances of the experiment - unquestionably about correct; so that the one originally observed of = -0.25 would have been entirely misleading.

6. Criticism of Prevelent Working Methods.

So far, our illustration of systematic deviation has been confined to instances taken from personal experience. But it might perhaps be thought that other workers avoid such perversions of fact by the simpler method of common sense. Unforutnately, such does not seem to have been at all the case; not once, to the best of my knowledge, has any partial association between two psychological events been determined in such a way as to present any good evidential value - these are strong terms, but I think, hardly exaggerated.

Psychologists, with scarcely an exception, never seem to have become acquainted with the brilliant work being carried on since 1886 by the Galton-Pearson school. The consequence has been that they do not even attain to the first fundamental requisite of correlation, namely, a presise quantitative expression. Many have, indeed, taken great pains in the matter and have constructed arrays of complicated numerical tables; but when we succeed in orienting ourselves in the somewhat bewildering assemblage of figures, we generally find that they have omitted precisely the few facts which are essential, so that we cannot even work out the correlation for ourselves.

This lack of quantitaive expression entails far more than merely diminished exactitude. For in consequence, the expreimenters have been unable to estimate their own results at all carrectly, some have believed themselves to demonstrate an entire absence of correspondence, when the latter has really been quite considrrable; whereas others have presented to the public as a high correlation what has really been very small and often well within the limits of mere accidental coincidence; these limits they have had no means of determining, and moreover their data were usually obtained in such a way as to make it unnecessarily large.

Seeing, thus, that even the elementary requirements of good correlation work described in the first part of this article have been so generally deficient we cannot be surprised to find that the moreadvanced refinements of procedure discussed in the second part have been almost wholly unregarded; so that the final results are saturated and falsified with every description of impurity. In this respect, unfortunately, it is no longer possible to hold up even the Galton-Pearson school as a model to be imitated. The latter must now perform the very different office of saving us from detailed criticism of inferior work, by enabling us to form an opinion as to how ~~fat~~far the defect permeates and vitiates even the best existent correlational research.

As example, we will take Pearson's chief line of investigation, Collateral Heredity, at that point where it comes into closest contact with our own topic, Psychology. Since 1989 he has, with government sanction and assistance, been collecting a vast number of data as to the amount of correspondence existing between brothers. A perliminary calculation, based in each case upon 800 to 1,000 pairs, led, in 1910 to the publication of the following momentous results:

Coefficients of Collateral Heredity.
Collelation of Pairs of Brothers.

Physical Characters.

(Family Measurements)

Stature	0.5107
Forearm	0.4912
Span	0.5494
Eye-colour	0.5169
(School Observations)	
Cephalic index	0.4861
Hair Colour	0.5452
Health	0.5303
Mean	0.5171

Mental Characters.

(School Observations)

Intelligence	0.4559
Vivacity	0.4702
Conscientiousness	0.5929
Popularity	0.5044
Temper	0.5068
Self-consciousness	0.5915
Shyness	0.5281
Mean	0.5214

Dealing with the means for physical and mental characters, we are forced to the perfectly definite conclusion that the mental characters in man are inherited in precisely the same manner as the physical. Our mental and moral nature is, quite as much as out physical nature, the outcome of heredity factors,

Now let us consider how these coefficients of correlation will be affected by our "systematic deviations." To begin with there is the "Attenuation" by errors; since it evidently cannot be assumed that the schoolmasters' judgments as to conscientiousness, temper, etc. are absolutely infallible. On page 90, it has been shown that deviation from this source may be estimated by the following domula:

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'p'}}}$$

To ascertain $r_{p'p'}$ and $r_{p'q''}$, I am aware of no precise data ~~between~~ beyond that found in some experiments of my own, where the independent intellectual gradings for the same series of subjects correlated with one another on an average to the amount of -.64. As on other occasions very competent persons have estimated this to be as much as should be expected, as as intelligence is about the most easily gradable of all the mental qualities mentioned by Pearson, t there is so far no reason to suppose that his "great number of masters and mistresses" did on the whole any better. Hence, even if we could assume that the mistakes in estimating one brother were independent of the mistakes in estimating the other, then the true correlation would be about, not 0.5172,

but $\frac{0.5172}{\sqrt{6.64 \times 0.64}}$ = 0.81, an extent of difference that seriously modifies our impression of exactitude from all these coefficients to four places of decimals. When we further consider that each of these physical and

mental heredity can hardly be more than mere accidental coincidence.

Let us next proceed to irrelevant correlation, and take for our theme postnatal accidents connected on the one side with brotherhood and on the other with the mental qualities. Pearson's primary intention seems to have been to make his correlations as "universal" as possible and in one place he expressly mentions that education is among the causes contributory to variation. Hence, he is more than consistent, in that he forms his correlation without regard to the fact that the ~~correlation~~ correspondence between the brothers' "conscientiousness," "popularity," etc., must be in great measure due to their coming under the same home influences. But such a correlation can scarcely be accepted as scientifically valuable. Fo we do not really know anything precise about the assimilating effects of heredity, when our observed correspondence is perhaps chiefly due to the brothers having the same amount of hampers and pocket-money. Still less can we, then, fairly compare such results with that obtained from physical measurements, where common home life has little or no effect. The factor of post-natal accidents, therefore, cannot but be regarded as irrelevant, and consequently the coefficients of correlation must be taken as hopelessly "distorted."

But even consistence cannot be upheld throughout the matter. For though the effect of postnatal life has thus been admitted with regard to education at home, it has perforce been excluded as regards public education. For only those brothers have been compared together who are at the same school; the coefficients of correlation would certainly diminish if those also could be included who are living in a totally different manner, have gone to sea, etc., The correlations are therefore also illegitimately "dilated."

If this work of Pearson has thus been singled out for criticism, it is certainly from no desire to undervalue it. The above and any other systematic errors are eventually capable of adequate elimination, and this article has itself, it is hoped been of some use towards that purpose. Such correction will no doubt necessitate an immense amount of further investigation and labour, but in the end his results will acquire all their proper validity. My present object is only to gu rd against premature conclusions and to point out the urgent need os still further improving the existing methodics of correlation work, a method of investigation which he himself has so largely helped to create and by means of which he is carrying light into immense regions hitherto buired inthe obscurity of irresponsible speculation. The fundamental difference between his prodedure and that here recommended, is that he seeks large natural samples of any existing series sufficiently homogeneous to be treated mathematically; whereas here smaller samples are deemed sufficient, but they are required to be artificially selected, ordered, and corrected into full scientific significance. His methods are those of pure statistics those inculcated here may be more aptly termed "statisticoids."

7. Number of Cases Desirable for an Experiment.

This leads us to the important question, as to how many cases it it advisable to collect for a single series of experiments. In actual practice, the greatest diversity has been apparent in this respect; many have thought to sufficiently establish important correlations with less than ten experimental subjects while others habe thought it necessary to gather together at least over a thousand.

mental [illegible] [illegible] [illegible]

Let us next proceed [illegible] [illegible] these [illegible] [illegible] other with [illegible] to [illegible] mentioned [illegible] more [illegible] that [illegible] "peculiarity," etc., [illegible] home influences. [illegible] valuable. [illegible] of heredity, [illegible] having [illegible] fairly [illegible] correspond [illegible] there are, [illegible] of correlation [illegible]

[illegible] though [illegible] as [illegible] brothers [illegible] of correlation would [illegible] living [illegible] correlations are therefore [illegible]

[illegible] it [illegible] it is hoped [illegible] mechanistic [illegible] his results [illegible] get [illegible] further [illegible] method of [illegible] ation [illegible] commonly [illegible] agic [illegible] concede [illegible] [illegible] deemed [illegible] corrected [illegible] these [illegible]

7. [illegible]

[illegible] it is [illegible] sufficiently [illegible] subjects [illegible] other factors at least over a thousand.

Now, a series of experiments is a very limited extract, whose disposition is, nevertheless, to be accepted as a fiar sample of the whole immens remainder. Other things equal, then, the l rger the sample the greater its evidential value and the less chance of a mere occassional coincidence being mistaken for the permanent universal tendency.

This danger of accidental deviation has been discussed in the first part and there shown to be strictly meaaurable by the "probable error." We there saw, also, that this danger can never be entirely eliminated by any sample however large, so that it is necesary to accept some standard less rigorous than absolute certainty as sufficient for all practical purposes; usually, the danger of mere chance coincidence is considered to be inappreciable when a correlation is observed as much as five times greater than the probable error, seeing that mere chance would not produce this once in a thousand times. Hence, evidently, the accidental deviation depends, not only on the number of cases, but also on the largeness of the really existing correspondence; the more perfect the latter, the fewer the cases that will be required to demonstrate it conclusively; and this tendency is augmented by the fact that the probable error, besides varying inversely with "n", does so to a further extent with "r" (see formula). It was shown in the same part that the size of the probable error also varies accordingly to the method of calculation- and to such an xtent that twenty cases treated in one of the ways described furnishes as much certitude as 180 in another more usual way. If the common trifold classification be adopted, an even greater number is required to effect the same purpose; and if the correlation be not calculated quantitatively at all, but instead be presented in the customary fashion to the reader's general impression, then no number of cases whatever apprear sufficient to give reasonable guarantee of proof.

While this the number of subjects is not by any means the sole factor in diminishing even the accidental deviation it has no effect whatever upon the far more formidable systematic de iation except that it indirectly leads to an enormous augmentation thereof . When we are taking great pains to be able to show upon paper an imposing number of cases and a diminutive probable error, we are in the self same provess most likely introducing a systematic deviation twenty times greater.

From this we may gather that the number of cases should be determined by the simple principle, that the measurements to be aggregated together should have their error brought to the same general order of magnitude. An astronomical chronometer, with spring-detent excapement, is not the best travelling clock; nor is there any real advantage in gracing upon a milestone (as has actually been done by an infatuated mathematician!) the distance to the nearest village in metres to three decimal places. Now, the present stage of Correlational Psychology is one of pioneering; and, instead of a few unwieldy experiments, we require a large number of small ones carefully carried out under varied and well considered conditions. At the same time, however, the probable error must be kept down to limits at any rate small enough for the particular object of investigation to be proved. For such a purpose a probable error may at present be admitted without mush hesitation up to about 0.05; so that, by adopting the method of calcudation recommended, two to three dozen subjects should be sufficient for most purposes. The precision can always be augmented subsequently, by carrying our similar experiments under similar conditions ahd then taking averages. Only after a long preliminary exploration of this rougher sort, shall we be in a position to effectually utilize experiments designed and executed from the very beginning on a vast scale.