THE STANDARD ERRORS OF CHROMOSOME DISTANCES AND COINCIDENCE[1]

H. J. MULLER AND JESSIE M. JACOBS-MULLER
University of Texas, Austin, Texas

Received March 28, 1925

TABLE OF CONTENTS

THE PROBLEM

Many problems in linkage require the comparison of two or more values obtained under different genetic or environic conditions, with the object of determining whether or not the observed differences between these values are "significant." By the term "significant difference" is here meant one of such size that it would be improbable for it to have arisen solely as a result of the random sampling of identical germinal material. For the purpose of such comparisons, then, it is necessary first to know the size of the deviations which random sampling by itself would be likely to cause. This is gauged by means of the "standard error," it being ordinarily true that deviations greater than two or three (according to the standard of certainty) times the standard error are very improbable, as a mere result of random sampling. (The use of the so-called "probable error" merely involves the standard error in somewhat different guise, as the former is ordinarily obtained by multiplying the latter by .6745,—a rather superfluous procedure except in special cases).

It has thus become almost axiomatic, for rigorous workers, that in order to be sure of their ground in the interpretation of their results they must have an idea of the standard errors of the values with which they deal. It is true that often the differences are so obviously decisive

[1] Department of Zoölogy, UNIVERSITY OF TEXAS, Contribution No. 194.

that great refinements are not necessary, and yet, unless some estimates are made of the "errors" involved, there will be occasions, not infrequent, when the investigator will either run the risk of being led into some serious misinterpretation, or else will fail to reap the full meaning from his results.

The standard error of the simple proportion of "crossovers," or, more accurately, of separations, between two pairs of genes,—as well as the error of a chromosome distance (in $\frac{\text{units}}{100}$) so short that it includes no double or multiple crossovers,—is well known, being determined by the familiar formula $\epsilon_p = \sqrt{\frac{p(1-p)}{n}}$, where ϵ_p is the standard error of the proportion of separations or crossovers, p, and n is the total number of individuals counted. (When p represents percents rather than proportions the formula is $\epsilon_p = \sqrt{\frac{p(100-p)}{n}}$.) But the standard errors of longer map lengths, involving double crossovers, and of the index of double-crossover frequency itself,—coincidence,—have not hitherto been worked out. As these are values just as important, in their way, and as frequently used in theoretical work, as the simple crossover values, it is essential that formulae be available for calculating their standard errors also.

STANDARD ERROR OF A MAP LENGTH

Let us consider first the standard error of a chromosome map, or a section of a map long enough to include double, etc., crossovers, based on a count involving simultaneously all the loci dealt with. The map length is the sum of the percent of crossovers in each of the separate regions; this is evidently the same as (100 times) the quotient formed by dividing the total number of individuals counted into the number of *crossings over* (as distinguished from crossovers,—each double crossover containing 2 crossings over, each triple crossover 3 crossings over, etc.). Thus, the map length of the regions considered is really (100 times) the mean value of the number of crossings over per individual in these regions. Now the standard error of any mean value (m) is equal to the standard deviation of the values of the individuals that go to make up the mean, divided by the square root of the number of such individuals $\left(\epsilon_m = \frac{\sigma}{\sqrt{n}}\right)$

In a given set of data, σ, the standard deviation of the values of the individuals, may be determined by the usual process, which consists of getting the square of the deviation of each individual value from the mean value, averaging these squares, and extracting the square root of this average, thus, $\sigma = \sqrt{\frac{\Sigma(i-m)^2}{n}}$, where m is the mean value, in our case the $\frac{\text{map length}}{100}$ or mean number of crossings over per individual, and i the individual value or number of crossings over in any given individual. This can also be expressed in the form, $\sigma = \sqrt{\Sigma\left(\frac{i^2}{n}\right) - m^2}$. It follows that $\epsilon_m = \sqrt{\frac{\Sigma\left(\frac{i^2}{n}\right) - m^2}{n}}$.

All that now remains is to find the result of substituting for $\Sigma\left(\frac{i^2}{n}\right)$ in the above formula the values derived from the data. Let s be the proportion of single crossovers in the entire total; in the case of each single crossover the value of i, and i^2, is 1, and the sum of the values $\frac{i^2}{n}$ is therefore s. Let d be the proportion of double crossovers; since each double crossover has a value of 2 for i, and of 4 for i^2, the sum $\left(\frac{i^2}{n}\right)$ for these is $4d$. Similarly, let t be the proportion of triple crossovers, the sum of $\left(\frac{i^2}{n}\right)$ for the triples being $9t$; for the quadruples it is $16q$, and so on. Then, the entire sum, $\Sigma\left(\frac{i^2}{n}\right)$, equals $s+4d+9t+16q\cdots$. But m, the map length, equals $s+2d+3t+4q\cdots$; hence, $\Sigma\left(\frac{i^2}{n}\right) = m+2d+6t+12q\cdots$ (or $m+2.1d+3.2t+4.3q\cdots$). Substituting this value of $\Sigma\left(\frac{i^2}{n}\right)$ in the formula for ϵ_m we have

$$\epsilon_m = \sqrt{\frac{(m+2.1d+3.2t+4.3q\cdots) - m^2}{n}},$$

or

$$\epsilon_m = \sqrt{\frac{m(1-m)+2.1d+3.2t+4.3q\cdots}{n}} \qquad (1)$$

It is evident that where there are no double or multiple crossovers to be considered, this expression reduces to the familiar $\sqrt{\frac{p(1-p)}{n}}$ for the standard error, ϵ_p, of the proportion of separations. The above value for ϵ_m, of course, applies to the mean *proportion* of crossings over, or $\frac{\text{map length}}{100}$, so that the standard error of the map length itself, when the latter is expressed in "chromosome units" or percents of crossings over, rather than in proportions, has a value 100 times that of the above expression. When, therefore, m represents the number of units of map length, and d, t, q, etc., represent the percents rather than the proportions of double, triple, quadruple, etc., crossovers, we have instead the relation

$$\epsilon_m = \sqrt{\frac{m(100-m)+200d+600t+1200q\cdots}{n}} \quad \ldots \quad (1a)$$

The above formulae can also be arrived at by considering the map as the sum of the various component distances,—a, b, c,. . . —and applying the equation for the standard error of a sum,

$$\epsilon_s = \sqrt{\epsilon_a^2+\epsilon_b^2+\epsilon_c^2\cdots\cdots+2r_{ab}\epsilon_a\epsilon_b+2r_{ac}\epsilon_a\epsilon_c\cdots\cdots+2r_{bc}\epsilon_b\epsilon_c\cdots\cdots},$$

where ϵ_a, ϵ_b, ϵ_c, etc., are the standard errors of a, b, c, etc., obtained by the $\sqrt{\frac{p(1-p)}{n}}$ formula, and r_{ab}, r_{ac}, r_{bc}, etc., are the correlations between a and b, a and c, b and c, etc., respectively. These correlations are obtained by the formula $r_{p_1p_2} = \frac{d-p_1p_2}{\sqrt{p_1p_2(1-p_1)(1-p_2)}}$, where d is the proportion of double crossovers, and p_1 and p_2 are the proportions of crossovers in the two respective regions considered.

Of course, the formulae given do not take into account possible errors due to the existence of unobserved double crossovers both of whose loci of crossing over lie between two "adjacent" genes, i. e., within the limits of a region indivisible in the experiment; such errors are caused by the conditions of the experiment, whereas the errors given by the formulae are merely those which would be caused by random sampling under these experimental conditions. Furthermore, the formulae do not take into account variations due to determinate causes other than sampling, such as genetic, "developmental," or environic circumstances that influence either crossing over, or the viability of different classes of offspring. As

such sources of variation are seldom absent except where the strictest attention has been given to genetic homogeneity of the parental material, and identity of age, and when the various experiments have been performed simultaneously, with the same food, etc., it would seem a supererogation to develop the formulae for the error of map length due to random sampling further at present, so as to include the errors of composite maps, formed by the combination of the results of wholly different experiments, involving different genes. The methods for determining the "most probable" value of the map from a combination of experiments with different loci have been worked out by FISHER (1922) and by KELLEY (1923), but the standard error of such a "most probable" map can only be estimated roughly, after numerous experiments have given a basis for judging the usual amount of variation due to "determinate" causes, among the results, for identical loci, of experiments involving different subsidiary loci and different environic conditions.

STANDARD ERROR OF COINCIDENCE

Coincidence is the ratio of the proportion of double crossovers (d) which actually occur in two regions (of "lengths" a and b in $\frac{\text{"units"}}{100}$), to the proportion of double crossovers which would occur there if crossings over in the two regions were independent of one another (the latter value being evidently ab). Thus the value of the coincidence ratio, c, is given by the formula $c=\frac{d}{ab}$.

In calculating the standard error of this ratio we may treat it as the quotient of two proportions, p_1 and p_0, where $p_1=\frac{d}{a}$ and $p_0=\frac{b}{1}$. Then, $c=\frac{p_1}{p_0}$. The numerator, p_1, is the proportion of crossovers in region b which occurs among the an cases having crossing over in region a, and the denominator, p_0, is the proportion of crossovers in region b which occurs in the entire total of n individuals. Thus, we are enabled to use for the standard error of coincidence, the well known close-approximation formula for the standard error of a quotient, which may be stated as follows:

$$\left(\frac{\epsilon_{p_1}}{p_0}\right)=\frac{p_1}{p_0}\sqrt{\left(\frac{\epsilon_{p_0}}{p_0}\right)^2-\left(\frac{\epsilon_{p_1}}{p_1}\right)^2-2r_{p_0p_1}\frac{\epsilon_{p_0}}{p_0}\frac{\epsilon_{p_1}}{p_1}}$$

(where ϵ represents the standard error, and r the correlation, of the values

given in the respective subscripts). To solve this expression for the present case we must find the values of ϵ_{p_0}, ϵ_{p_1} and $r_{p_0p_1}$, and substitute them in the equation given.

As p_0 is merely a proportion (b) of a fixed total (n) its standard error in random sampling is accurately given by the formula $\epsilon_{p_0} = \sqrt{\frac{p_0(1-p_0)}{n}}$. Somewhat similarly, p_1 is a proportion of the "total" n and its standard error, ϵ_{p_1}, may be taken as $\sqrt{\frac{p_1(1-p_1)}{an}}$. As will be explained later, however, the latter expression is only an approximation to ϵ_{p_1}, since the observed value of an is itself subject to variation. In obtaining the value of $r_{p_0p_1}$, it should be noted that the proportion p_0 is gotten by the inclusion of the individuals that go to form p_1 (the double crossovers), with others (single crossovers in region b), to form a proportion of a larger total (n). The formula for the correlation of two such proportions,—one based on individuals that are also included in the other,—is $r = \frac{n_1}{n_0}\frac{\epsilon_{p_1}}{\epsilon_{p_1}}$, where n_1 is the smaller total, in this case an, out of which p_1 is obtained, and n_0 is the more inclusive total, in which p_0 occurs. In the present case, then, $r_{p_0p_1} = \frac{an\epsilon_{p_1}}{n\epsilon_{p_0}} = \frac{a\epsilon_{p_0}}{\epsilon_{p_1}}$

Substituting, now, the above values of ϵ_{p_0}, ϵ_{p_1}, and $r_{p_0p_1}$ in the formula for the standard error of a quotient and simplifying each term, we obtain $\epsilon_c = c\sqrt{\frac{1-p_0}{p_0n} + \frac{1-p_1}{p_1an} - 2\left(\frac{1-p_1}{p_0n}\right)}$. Next, substituting for p_0 and p_1 their values b and $\frac{d}{a}$:

$$\epsilon_c = c\sqrt{\frac{1-b}{bn} + \frac{a-d}{adn} - \frac{2(a-d)}{abn}}$$

Reducing to common denominator, $\epsilon_c = c\sqrt{\frac{ad(1-b)+b(a-d)-2d(a-d)}{abdn}}$

Simplifying the numerator and rearranging terms,

$$\epsilon_c = c\sqrt{\frac{ab-d(a+b+ab-2d)}{abdn}}$$

Substituting for $\frac{d}{ab}$ its value c, and for $(a-d)$ and $(b-d)$ the symbols a_s and b_s, respectively, signifying the proportion of *single* crossovers in regions a and b, we have, finally,

$$\epsilon_c = c\sqrt{\frac{1-c(a_s+b_s+ab)}{dn}} \quad \text{(approximately)} \quad \cdots \quad (2)$$

For much work it will be found sufficiently accurate to use in place of this a rougher approximation, derivable from it, as follows:

$$\epsilon_c = c\sqrt{\frac{1-cm}{D}} \quad \text{(approximately)}, \quad \cdots \quad (2a)$$

where m is the $\frac{\text{map length}}{100}$ of the regions in question, i. e., $m=a+b$, and $D=dn$, the absolute number of double crossovers.

Another form of formula (2), sometimes more convenient in practice, may be obtained from the expression just preceding (2) by dividing the terms of the numerator into the denominator. We then have:

$$\epsilon_c = c\sqrt{\frac{1}{dn} - \frac{1}{bn} - \frac{1}{an} - \frac{1}{n} + \frac{2c}{n}}.$$

Denoting dn, bn and an by D, B and A, which are the absolute numbers rather than the proportions of double crossovers and of crossovers in regions b and a, respectively, we have:

$$\epsilon_c = c\sqrt{\frac{2c-1}{n} - \frac{1}{A} - \frac{1}{B} + \frac{1}{D}} \quad \text{(approximately)} \quad \cdots \quad (2b)$$

As all these formulae are symmetrical with respect to a and b (that is, the latter may be interchanged without altering the final value) it is evident that the use of $\frac{d}{b}$ as p_1 and of a as p_0 in working out the result would have led to the same expression. Nevertheless, as mentioned previously, even formulae (2) and (2b) are not strictly accurate, first, because the formula used for the standard error of a quotient is only an approximation, though a very close one, and second, because the formula

for ϵ_{p_1} is approximate, since the observed values representing the "total" an are variable.

Where the total out of which a proportion is taken is variable, strict accuracy usually demands the use of H, the harmonic mean of these totals, rather than the arithmetic mean (in our case an), in the place of n in the formula $\epsilon_p = \sqrt{\frac{p(1-p)}{n}}$. Where the actual values of the totals are not available for determining H, the latter may usually be calculated from the arithmetic mean, n, by the approximation formula, $H = n\left(1 - \frac{\sigma_n^2}{n^2}\right)$. Substituting, for our case, an for n and $\sqrt{\frac{an(1-an)}{n}}$ for σ_n, we have $H = an + a - 1$. If this is used in place of an in the formula for ϵ_{p_1}, and the formula for ϵ_c worked out by steps similar to those taken previously, we obtain:

$$\epsilon_c = c\sqrt{\frac{2c-1}{n} + \frac{a}{an+a-1}\left(\frac{1}{d} - \frac{1}{b} - \frac{1}{a} + \frac{1-a}{abn}\right)} \quad \text{(approximately)} \quad (3)$$

This is obviously unsymmetrical with respect to a and b, due to the fact that the formula for H used was only an approximation. In fact, it can be shown that the difference between the value of this expression and that obtained when a and b are interchanged would not infrequently be greater, in cases of the sort dealt with experimentally, than the difference between one of them and the original formula (2). Doubtless a better approximation could be obtained by using the mean of a and b rather than a in the unsymmetrical portions of formula (3). This works out as follows:

$$\epsilon_c = c\sqrt{\frac{2c-1}{n} + \frac{a+b}{(n+1)(a+b)-2}\left(\frac{1}{d} - \frac{1}{a} - \frac{1}{b} + \frac{2-a-b}{2abn}\right)} \quad \text{(approximately)} \quad (3a)$$

The error caused in actual problems by the use of the arithmetic rather than the approximate harmonic mean for an is, however, never more than a few percent of the value of ϵ_c. Such an amount is usually of negligible consequence when standard errors are dealt with, for the latter are ordinarily used for determining in round numbers (or numbers of the accuracy of 2.5) the multiple which a given deviation is of the standard deviation. Hence, there would seldom be reason for employing the unwieldy formula (3) or (3a) rather than (2) or (2a).

In strictest accuracy, ϵ_c is indeterminate, because in reality ϵ_{p_1} is indeterminate. The real harmonic mean, H, of observed values of an, must always be 0, since in a practically negligible proportion of samples an will be 0, and the harmonic mean of any series including 0 must always be 0 also; this in turn would make $\epsilon_{p_1}=0$. It is more correct, however, to consider the deviations of p_1 itself, since the root mean square of these really give ϵ_{p_1}. If we do this, we find that when $an=0$, since the deviation of d also $=0$, the proportion p, being $\frac{d}{an}$, or $\frac{0}{0}$, is indeterminate, and its deviation is therefore indeterminate also. This causes ϵ_{p_1} to be indeterminate, and consequently ϵ_c, even though a sample in which an was 0 would not occur, in ordinary work, once in a billion times (an would be zero about once in 200,000,000,000 times if a represented 5 units and n a total of 500 individuals).

If, then, we use the term standard error in the most rigorously exact sense we see that in the case of coincidence its value cannot be found, and does not, in fact, exist, as a definite quantity. Nevertheless, we can continue to speak of it, and to use one of the above formulae for ϵ_c in our work, and these values will have a practical meaning similar to that of the standard error of other quantities, inasmuch as a random deviation of a certain number of times this ϵ_c will have about the same amount of probability as a random deviation of another quantity which is the same number of times its standard deviation. And the "probable error," as in other cases, will here too be about .6745 times the value taken as representing standard error, provided dn is a reasonably large number. In the case of coincidence, in fact, the "probable error" is really a value of more definite meaning than the standard error, being determinate, and independent of the indeterminate value of $\frac{d}{an}$ for the cases when $an=0$. Usually, however, it would be necessary first to determine ϵ_c as above, before the probable error could be found.

It should also be noted that although ϵ_c, strictly speaking, is indeterminate, the range of indetermination is very small, since for all ordinary values of a and n used the proportion of samples in which $an=0$ is exceedingly minute, and in each of the latter samples, even though $\frac{d}{an}=\frac{0}{0}$, this ratio can never be greater than 1 (nor less than 0), as d can never exceed an. The standard error, then, which involves the sum of

numerous determinate quantities and these few indeterminate quantities of limited value, becomes very narrowly fixed. Theoretically, limits could be assigned to ϵ_c and it would be found that the difference between these limits is utterly negligible compared with the size of ϵ_c itself, or compared with the difference between either of them and the practically adequate value given by approximation formula (2). Thus, the question of the more exact determination of ϵ_c ceases to be of practical moment.

THE USE OF THE STANDARD-ERROR FORMULAE

Errors of observed versus true values

Every one of the formulae mentioned so far, including those for crossovers, map length and coincidence, has given the standard deviation to be observed in a large collection of random samples if the true values (for proportion of crossovers, double crossovers, coincidence, etc.), that is, the values which would be found in an indefinitely large sample of the same material, were those used in the formula. Values resulting from random sampling of this material that deviate from the "true" value by more than two or three times this standard error may then be taken as improbable, since they can be shown to occur infrequently, and when such values are found it is therefore considered probable that they were drawn from material with a true value different from that assumed.

In practice, however, the question usually to be answered is not the above,—what the observed values may be which have the greatest reasonable deviation from a certain assumed true value,—but the converse, that is, what the true values could be which would have as their greatest "reasonable" deviant a certain observed value. We can not answer this question precisely by the simple use of the preceding formulae, since the standard error, given in the formula, of a true value equal to that observed, is not precisely the same as the standard error of a true value differing from that observed by plus or minus two or three times the latter standard error itself. However, the values of these errors are usually sufficiently alike that one may be used in place of the other without serious danger of an erroneous conclusion, unless the absolute number of one or more of the variants involved is extremely small, and it has accordingly been the practice to use such formulae as the above for finding the limits of the true values which observed values may represent, by adding to and subtracting from the latter 2 or 3 times the error given by the formulae. It is as legitimate to do this in the case of the map and co-

incidence formulae as in the case of the other formulae where this is commonly done.

When greater accuracy is desired, it is customary to use a rather cumbersome method of approximation. In this method the observed value is first assumed to be true, and by the aid of a formula like one of those given above, the plus and minus limits of the "reasonably possible" observed values (differing from the former by 2 or 3 times the standard error) are calculated. These are then assumed to be true, and their standard errors are calculated by the same formula. Deviations from the observed value of two or three times these, in the plus or minus direction, respectively, now give the true values to a second approximation. The same process may be repeated as many times as necessary, until the desired degree of accuracy is attained. In the case of coincidence, this procedure would be considerably more difficult and intricate than would appear from the above outline, since the standard error of any true or assumedly true value of coincidence is a function not only of the coincidence itself, and the total number counted, but also of the different classes of crossovers, the values of which vary in partial independence of one another. Just how to take all these variations into account simultaneously is not at present clear.

When we are dealing with the simple proportion of crossovers, however, or any other simple proportion (such as of non-disjunctional exceptions, mutations, etc.), the above approximation method may be replaced by a more direct and exact procedure. Let p_0 be the observed value of the proportion and p_1 and p_2 the respective larger and smaller possible true values which differ from p_0 by a certain number of times, say a times, their own standard error. Then we have $p_1 - p_0 = a\sqrt{\frac{p_1(1-p_1)}{n}}$, and $p_0 - p_2 = a\sqrt{\frac{p_2(1-p_2)}{n}}$. If we solve these equations for p_1 and p_2, respectively, we find that $p_1 = \frac{2np_0 + a^2 + a\sqrt{4np_0(1-p_0)+a^2}}{2(n+a^2)}$ and p_2 equals an expression which is the same as the above except that a minus sign occurs before the term containing the radical. Thus, if we let p_t represent either extreme possible true value, we have

$$p_t = \frac{2np_0 + a^2 \pm a\sqrt{4np_0(1-p_0)+a^2}}{2(n+a^2)} \quad \ldots\ldots \quad (4)$$

where p_0 is the observed value and a the number of times the standard deviation of the possible true value whereby the latter differs from the observed value. This formula does not seem to be well known, but, although somewhat lengthy, it is necessary where exactitude is sought, and it is especially important when pn is a rather small number.

Errors due to causes other than random sampling

A second point which must be kept in mind in the application of any of the formulae above discussed is that they give an idea of the size of such deviations as result from random sampling alone. A deviation greater than that thus indicated would not prove the effectiveness of a given factor or agency in influencing the value studied unless it could be shown that no other variation was possible in the experiment except that due to random sampling and to this agency. This is seldom the case in work on linkage, non-disjunction, and other genetic processes giving irregular ratios, and so the unmodified formulae of random sampling are only applicable in the comparison of experiments in which the strictest attention has been given to uniformity of genetic and other conditions in all respects except those the influence of which it is desired to determine (or those the amount of influence of which is definitely predictable). As GOWEN'S (1919) work shows, even in such cases there may be uncontrollable sources of variation making the deviations greater than in random sampling.

Wherever possible, then, statistical tests should be applied to the material, by getting the results of various samples taken under the (supposedly) same conditions, and determining whether or not the deviations of these samples from one another are greater than would be expected of purely random samples. The formula for this test is easy, since the deviations of the values in the samples from the general mean value, when squared, summed and averaged, so as to get the standard deviation of these values, should not differ significantly from the standard error to be expected of the average sample (determined by one of the above formulae, with the use of H, the harmonic mean, as the mean number per sample). Whether the resultant difference is significant may usually be found with sufficient accuracy by the use of the approximate formula for the standard error of a standard deviation, that is, $\frac{\epsilon}{\sqrt{2N}}$ in which ϵ is the calculated standard error of the samples and N is the number of samples. These methods apply alike to problems of map length, coin-

cidence, percent of crossovers, of non-disjunction, etc. Of course a satisfactory agreement of the observed deviations of the samples with the error expected from random sampling is not a sure proof that other sources of variation may not be at work, but a contrary result,—a significant disagreement,—does prove that the unmodified random sampling formulae do not apply.

If, by reason of tests like the above or on account of *a priori* considerations, it is concluded that the random sampling rules are insufficient, there may remain another mode of procedure for determining whether a given condition or set of conditions is exerting a significant influence upon the genetic phenomenon studied, or for determining the amount of such influence. This, however, like the above test, requires that a considerable number of separate samples have been recorded, preferably in both (or all) of the series to be contrasted. The standard deviation of the values of the separate samples from the general mean value is then calculated for all the series taken together, by the same method as used in the tests discussed above, and this standard deviation, σ, divided by the square root of the number of samples (N_1) comprised in a given series, will give the standard error $\left(\frac{\sigma}{\sqrt{N_1}}\right)$ allowed for the mean of this entire series, provided the special controlled agency which differentiates one series from another is ineffective in influencing the genetic process studied. Similarly, $\frac{\sigma}{\sqrt{N_2}}$, the standard error of the mean of the entire second series, composed of N_2 samples, may be obtained. These two quantities can now be used in the familiar formula for the standard error of a difference, where there is no correlation, $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2}$, to determine in this case the standard error for the difference of the means of the two entire series. If, then, the actual difference between these means is more than two or three times the latter standard error, it may be concluded that the agency studied has been effective.

This final conclusion will be valid even if there were numerous other agents affecting the genetic process studied, so long as there was no cause other than "chance" to lead these agents to act on the samples of one series rather than the other,—that is, if each sample, independently of every other, were as likely to come under the influence of one or more of these agents as every other sample. For the diversifying influences

of these extraneous factors, uncorrelated with the controlled factor, has been allowed for by taking the observed standard deviations of the samples rather than the errors calculated on the basis of random sampling. But, as before, while a significant difference between the means of the series will thus prove the effectiveness of an agent, the lack of such a difference will not categorically disprove the latter but will merely assign an upper limit to it.

The above method is theoretically applicable no matter whether the value studied be proportion of crossovers, of non-disjunction or other exceptions, map length, coincidence, or anything else. In the case of coincidence, however, since this requires such large numbers in a sample for a single good determination, it is often impracticable to secure large numbers of samples, but the work can usually be divided into a few samples, at least, so that some estimate can be obtained of the amount of variability due to all "extraneous" causes combined. Thus, an idea of the upper limit of such variability may be formed, by which the significance of the differences observed in different series may be gauged.

Where, however, the coincidence values to be compared concern different regions all of which were studied in the same counts, the formulae of random sampling (2, 2a, 2b) are accurately applicable, provided the approximate equality of contrary classes shows that the effects of differential viability are negligible. For in such a case identical genetic, developmental and other environic factors were acting in the formation of the different gametic coincidence ratios, and the only possible sources of difference in the observed coincidences, aside from the effects of random sampling, are those inherent in the behavior of the different regions concerned and selective agents which may cause the adult ratios to differ from the gametic ones.

Comparisons of values

The formula for the standard error of a difference, of course, applies both in cases like those previously discussed, where the differences between means are dealt with, and also in all cases of purely random sampling. Since this process of getting the root sum of two squares must often be performed repeatedly, it is convenient to use a geometric scheme for making the computation (just as in multiplications and divisions we may use the slide rule). For the present calculation the authors find that if a sheet of coordinate paper be used, with the lines numbered by tens, both down and across, and another numbered piece of the paper, in the form of a strip,

be taken as a ruler, sufficient accuracy is attained by reading the distance subtended on the ruler when this is placed diagonally from a point on the upper edge having a numerical value equal to one of the standard deviations to a point along the left vertical edge having a value equal to the other standard deviation. This method, which obviously depends upon a hypotenuse being equal to the root sum of the squares of the sides, has been found to save considerable time and to be far quicker for this purpose than the slide rule.

Not merely the significance of a difference, but also the limits allowed for the intensity or degree of effect produced, are determined by the formula for the standard error of a difference. Intensities of effect are, however, expressed more intelligibly, and are more readily dealt with, by means of the quotients than by the differences of the values found in different series. The formula for the standard error of quotients of uncorrelated quantities in general has been mentioned in the section on coincidence. For the handling of these quotients the reader may be referred to the examples treated in the account of the effect of X rays upon crossing over in Drosophila autosomes (MULLER 1925).

SUMMARY

1. The formula is given (formulae 1 and 1a) for the standard deviation which would result from random sampling in the case of a chromosome map, or section of a map, the loci involved in which are followed simultaneously.

2. It is shown that the standard error of coincidence is not finally determinate, but that ordinarily its value is very narrowly limited. Formulae (2, 2a, 2b, 3, 3a) are presented, that give with various degrees of approximation the standard deviation of coincidence which would occur in random sampling.

3. Cautions to be observed in the use of these and other formulae for the standard deviations caused by random sampling are pointed out. Methods are reviewed for determining the significance of results in case other sources of variation besides random sampling and the possible influence of the factors to be studied unavoidably enter into the experiment.

4. The formula (4) is given for determining the maximum and minimum "possible" true values of a proportion of crossovers or of other genetic types which might, in random sampling, have been represented by a given observed value. This gives results somewhat different from those obtained by the formula in common use for this purpose.

LITERATURE CITED

FISHER, R. A., 1922 The systematic location of genes by means of crossover observations. Amer. Nat. **56:** 406–411.

GOWEN, J. W., 1919 A biometrical study of crossing over. On the mechanism of crossing over in the third chromosome of *Drosophila melanogaster*. Genetics **4:** 205–251.

KELLEY, T. L., 1923 Statistical method. 390 pp. New York: Macmillan Co. (See p. 321–324.)

MULLER, H. J., 1925 The regionally differential effect of X rays on crossing over in autosomes of Drosophila. Genetics **10:** 470–507.