THE ANALYSIS OF VARIANCE WHEN EXPERIMENTAL ERRORS FOLLOW THE POISSON OR BINOMIAL LAWS

By W. G. Cochran

1. Introduction. The use of transformations has recently been discussed by several writers [1], [2], [3], [4], in applying the analysis of variance to experimental data where there is reason to suspect that the experimental errors are not normally distributed. Two types of transformations appear to be coming into fairly common use: $\sqrt{x}$ and $\sin^{-1}\sqrt{x}$. The former is considered appropriate where the data are small integers whose experimental errors follow the Poisson law, while the latter applies to fractions or percentages derived from the ratio of two small integers, where the experimental errors follow the binomial frequency distribution. In each case the object of the transformation is to put the data on a scale in which the experimental variance is approximately the same on all plots, so that all plots may be used in estimating the standard error of any treatment comparison. The extent to which these transformations are likely to succeed in so doing has been examined by Bartlett [2]. The object of the present paper is to discuss the theoretical basis for these transformations in more detail, and in particular to examine their relation to a more exact analysis.

2. Experimental variation of the Poisson type. The first step in an exact statistical analysis of the results of any field experiment, is to specify in mathematical terms (1) how the expected values on each plot are obtained in terms of unknown parameters representing the treatment and block (or row and column) effects (2) how the observed values on the plots vary about the expected values. In this section, the variation is assumed to follow the Poisson law.

The specification of the expected values requires some consideration. In the standard theory of the analysis of variance, treatment and block (or row and column) effects are assumed to be additive. In the case of a Latin square, for example, the expected yield m_i of the ith plot, which receives the tth treatment and occurs in the rth row and the cth column is written

$$m_i = G + T_t + R_r + C_c \tag{1}$$

where G is a parameter representing the average level of yield in the experiment, and T_t, R_r and C_c represent the respective effects of the treatment, row and column to which the plot corresponds. Since the T, R and C constants are required only to measure differences between different treatments, rows and columns, we may put

$$\sum_t T_t = \sum_r R_r = \sum_c C_c = 0. \tag{2}$$

If the experimental errors are normally and independently distributed with equal variance, this specification leads to very simple equations of estimation for the unknown parameters, the maximum likelihood estimate of T_t, for example, being the difference between the mean yield of all plots receiving that treatment and the general mean. In addition to its simplicity, this type of prediction formula is fairly suitable for general use, because it gives a good approximation to most types of law which might be envisaged, provided that row and column differences are small in relation to the mean yield. However, in considering an exact analysis with Poisson variation, the prediction formula is assumed chosen, without reference to computational simplicity, as being the most suitable to describe the combined actions of treatment and soil effects.

The probability of obtaining a given set of plot yields x_i with expectations m_i may be written

$$\prod_i \frac{e^{-m_i} m_i^{x_i}}{x_i!}.$$

Thus L, the logarithm of the likelihood, is given by

$$L = \sum_i (x_i \log m_i - m_i) - \sum_i \log x_i! \tag{3}$$

Hence the maximum likelihood equation of estimation for any parameter θ assumes the form

$$\Sigma \frac{(x_i - m_i)}{m_i} \frac{\partial m_i}{\partial \theta} = 0 \tag{4}$$

where the summation extends over all plots whose expectations involve θ. The function $\frac{\partial m_i}{\partial \theta}$ will usually involve a number of parameters. Since the specification of row, column and treatment effects in a 6 x 6 Latin square requires 16 independent parameters, the solution of these equations may be expected to be laborious, though it may be shortened by the intelligent use of iterative methods. The problem of obtaining exact tests of significance is also difficult. The method of maximum likelihood provides estimates of the variances and covariances of the treatment constants, which under certain conditions can be assumed to be normally distributed if there is sufficient replication, but this can hardly be considered an exact "small sample" solution.

These remarks show that the exact solution is somewhat too complicated for frequent use. The difficulty arises principally because the typical equation of estimation consists of a *weighted* sum of the deviations of the observed from the expected values, the weights being $\frac{1}{m_i} \frac{\partial m_i}{\partial \theta}$. The factor $\frac{1}{m_i}$ was introduced into the weight by the Poisson variation of the experimental errors, and must be retained in any theory which claims to apply to Poisson variation. It is, however, worth considering whether some simplification cannot be introduced into

the equations by assuming some particular form for the prediction formula. This line of approach seems promising when one considers the simplification introduced into the "normal theory" case by assuming the prediction formula to be linear.

For Poisson variation, the linear law does not appear to be particularly suitable, since it may give negative expectations on some plots (as happens in the numerical example considered in the next section). Further, while $\frac{\partial m_i}{\partial \theta}$ becomes a constant, the factor $\frac{1}{m_i}$ remains in the weight.

The entire weight can be made constant by assuming a linear prediction formula in the square roots and transforming the data to square roots. For a Latin square, this prediction formula is written

$$\sqrt{m_i} = \alpha_i = G + T_t + R_r + C_c, \tag{5}$$

where

$$\sum_t T_t = \sum_r R_r = \sum_c C_c = 0. \tag{6}$$

To find the maximum value of (3) subject to the restrictions (6), we may use the method of undetermined multipliers, maximizing

$$L + \lambda(\sum_t T_t) + \mu(\sum_r R_r) + \nu(\sum_c C_c). \tag{7}$$

The equation of estimation for a typical treatment constant T_t becomes

$$\Sigma\left(\frac{x_i - m_i}{m_i}\right)\frac{dm_i}{d\alpha_i}\frac{\partial \alpha_i}{\partial T_t} + \lambda = 0, \quad \text{i.e.,} \quad \Sigma\frac{2(x_i - m_i)}{\sqrt{m_i}} + \lambda = 0, \tag{8}$$

the summation being extended over all plots receiving the treatment. If $a_i = \sqrt{x_i}$, then by Taylor's theorem

$$x_i - m_i = (a_i - \alpha_i)\frac{dm_i}{d\alpha_i} + \frac{1}{2!}(a_i - \alpha_i)^2\frac{d^2 m_i}{d\alpha_i^2} + \cdots \tag{9}$$

If m_i is reasonably large, only the first term on the right-hand side need be retained. When m_i is small, we may use, instead of the exact square root, a quantity a_i' defined so that

$$x_i - m_i = (a_i' - \alpha_i)\frac{dm_i}{d\alpha_i} = 2\sqrt{m_i}(a_i' - \alpha_i). \tag{10}$$

Thus if the analysis is performed on the quantities a_i' instead of on the original data, equation (8) becomes

$$\sum_{T_t} 4(a_i' - \alpha_i) + \lambda = 0. \tag{11}$$

On substituting the expectations for α_i from (5), and using (6), we obtain

$$\sum_{T_t} 4(a_i' - G - T_t) + \lambda = 0. \tag{12}$$

The corresponding equation for G is

$$\sum_{i} 4(a_i' - G) = 0, \tag{13}$$

so that G is the general mean of the quantities a'. By adding equations (12) over all treatments, and comparing the total with (13), we find $\lambda = 0$ Hence T_t is the difference between the mean yield of a' over all plots receiving T_t and the general mean of a'. In this scale the simplicity of the "normal theory" equations has apparently been recovered. Actually, the quantities a' are not known exactly, since

$$a' = \alpha + \frac{(x - m)}{2\sqrt{m}} = \frac{1}{2}\left(\alpha + \frac{x}{\alpha}\right) \tag{14}$$

where α is the expected value of $\sqrt{x}$. However, this process provides a means of successively approximating the maximum likelihood solution, by choosing first approximations to the quantities α, constructing the a''s, solving for the unknown constants and hence obtaining second approximations to the expected values. The close relation of a' to $\sqrt{x}$ is seen by remembering one of the common rules for finding square roots. This consists in guessing an approximate root (α), dividing x by the approximate root, and taking the mean of the approximate root (α) and the resulting quotient (x/α).

The suitability of the linear prediction formula in square roots must be considered in any example in which the above analysis is being employed. The law is intermediate in its effects between the linear law and the product law in the original data. My experience is that it is fairly satisfactory for general use, (cf. [2], p. 72) An exception may occur when it is desired to test the interaction between two treatments, both of which produce large effects. In this case the definition chosen for absence of interaction may not coincide at all closely with the definition implied in using the linear law in square roots. An example of this case was given in a previous paper [1].

In this connection it should be noted that an approximate "goodness of fit" test may be obtained of the validity of the assumptions made. Since the quantities a_i' enter into the equations of estimation with weight 4, the quantity $4\sum_i (a_i' - \alpha_i)^2$ is distributed approximately as χ^2 with the number of degrees of freedom in the error term of the analysis of variance. Some idea of the closeness of the approximation may be gathered by considering the simplest case in which only the mean yield is being estimated. In this case the observed values x are assumed to be drawn from the same Poisson distribution, and the sufficient statistic for the mean G is known to be $\Sigma(x_i)/n$. Since, however, the

prediction formula is here the same in square roots as in the original scale, and since the maximum likelihood solution is invariant to change of scale, the mean value α of a' must be *exactly* $\sqrt{\Sigma(x)/n}$, as the reader may verify by working any particular example. Thus $\Sigma 4(a' - \alpha)^2$ is found to be $\Sigma(x - \bar{x})^2/\bar{x}$, the usual χ^2 test for examining whether a set of values x may reasonably be assumed to come from the same Poisson distribution. By working out the exact distribution of $\Sigma(x_i - \bar{x})^2/\bar{x}$ in a number of cases [5], I previously expressed the opinion that this quantity followed the χ^2 distribution sufficiently closely for most practical uses, even for values of the mean as low as 2. This opinion has since been substantiated by Sukhatme, [6] who sampled this distribution for $m = 1, 2, 3, 4$, and 5.

A high value of χ^2 means either that the prediction formula is not satisfactory or that the experimental errors are higher than the Poisson distribution indicates, or that both causes are operating. These effects can sometimes be separated by examining whether the observed yields deviate from the expected yields in a systematic or a random manner. If the deviation is systematic, the prediction formula is probably unsatisfactory.

The type of approach used above resembles in many features the "exact" analysis for the probit transformation [7]. The principal difference is that in the case of probits the transformation is made to suit the *a priori* prediction formula, which postulates that the probits are a linear function of the dosage, or of the log (dosage). Thus with probits the equations of estimation still involve weights in the transformed scale. These do not seriously complicate the analysis, since only two parameters require to be estimated for a given poison. With, however, the much greater number of parameters usually involved in specifying the results of a field experiment, the attractiveness of a solution which does not involve weighting is greatly increased.

3. Numerical example of the square root transformation. A 5×5 Latin square experiment on the effects of different soil fumigants in controlling wireworms was selected as an example. The average number of wireworms per plot (total of four soil samples) was just under five. Previous studies [8], [9] have indicated that with small numbers per sample, the distribution of numbers of wireworms tends to follow the Poisson law.

The plan and yields are shown in Table I. The first two figures under the treatment symbols are the numbers of wireworms and their square roots respectively, the latter being regarded as first approximations to the values a'. Two of the plots receiving treatment K gave no wireworms. Since these plots are likely to be changed most in the transition from square roots to a', better approximations were estimated for them before proceeding with the calculations. The best simple approximations appeared to be obtained from the square roots of the means in the original units. For the plot in the second row and second column, the square roots of the row, column and treatment means in the original

TABLE I

Plan and number of wireworms per plot

	P	*O*	*N*	*K*	*M*	Mean
	3[1]	2	5	1	4	
	1.73[2]	1.41	2.24	1.00	2.00	1.676[2]
	1.76[3]	1.45	2.25	1.11	2.00	1.714[3]
	1.77[4]	1.46	2.25	1.10	2.00	1.716[4]
	M	*K*	*O*	*N*	*P*	
	6	0	6	4	4	
	2.45	(0.39)	2.45	2.00	2.00	1.858
	2.45	0.32	2.50	2.02	2.02	1.862
	2.46	0.32	2.49	2.02	2.02	1.862
	O	*M*	*K*	*P*	*N*	
	4	9	1	6	5	
	2.00	3.00	1.00	2.45	2.24	2.138
	2.10	3.09	1.00	2.47	2.25	2.182
	2.13	3.08	1.00	2.46	2.25	2.184
	N	*P*	*M*	*O*	*K*	
	17	8	8	9	0	
	4.12	2.83	2.83	3.00	(0.79)	2.714
	4.18	2.84	2.83	3.00	0.77	2.724
	4.17	2.84	2.83	3.00	0.77	2.722
	K	*N*	*P*	*M*	*O*	
	4	4	2	4	8	
	2.00	2.00	1.41	2.00	2.83	2.048
	2.14	2.02	1.49	2.04	2.92	2.122
	2.10	2.03	1.50	2.05	2.90	2.116
Mean	2.460[2]	1.926	1.986	2.090	1.972	2.087[2]
	2.526[3]	1.944	2.014	2.128	1.992	2.121[3]
	2.526[4]	1.946	2.014	2.126	1.988	

Treatment Means

K	*P*	*O*	*M*	*N*
1.036[2]	2.084	2.338	2.456	2.520
1.068[3]	2.116	2.394	2.482	2.544
1.058[4]	2.118	2.396	2.484	2.544

[1] Original numbers. [2] Square roots. [3] Second approximations. [4] Third approximations.

units are respectively 2.000, 2.145 and 1.095, and the square root of the general mean is 2.227. Hence

$$a' = \tfrac{1}{2}[2.000 + 2.145 + 1.095 - 2(2.227)] = 0.39.$$

The other zero value was similarly found to give $a' = 0.79$. The corresponding estimates from the means of the square roots were considerably too low, since the a' values tend to be higher than the square roots. The use of "missing plot" technique gave very poor approximations, because it ignores the fact that the plots in question had zero yields.

With the estimated values inserted, the row, column, and treatment means of the square roots are as shown in Table I. A second approximation to a' was calculated for each plot. For the plot in the first row and the first column, the expected yield is

$$\alpha = 1.676 + 2.460 + 2.084 - 2(2.087) = 2.046.$$

Hence $a' = \tfrac{1}{2}(2.046 + 3/2.046) = 1.76$. These values constitute the third set of figures in Table I. Theoretically, it is advisable to readjust the row, column, and treatment means after each new value of a' has been obtained, in order to secure rapid convergence. This is rather laborious in practice, and a complete set of new plot values was obtained before readjusting the means. The third approximations obtained by this method are shown in the fourth lines in Table I and are correct to two decimal places.

It is noteworthy how closely the square roots agree with the third approximations on all plots except those which originally gave zero yields. The differences between the second and third approximations are trivial.

The next step is to make a χ^2 test by means of the quantity $4\Sigma(a' - \alpha)^2$. From the manner in which the values α are constructed from the a''s, it follows that $\Sigma(a' - \alpha)^2$ is simply the error sum of squares in the conventional analysis of variance of the values a'. The analysis of variance of the third approximations is shown in Table II.

TABLE II

Analysis of variance of adjusted square roots

	Degrees of freedom	Sum of squares	Mean square
Rows	4	2.9815	
Columns	4	1.1190	
Treatments	4	7.5815	1.8954
Error	12	4.5970	0.3831

The value of χ^2 is $4 \times 4.597 = 18.39$, with 12 degrees of freedom, which is just about the 10 percent level. If the hypothesis is regarded as disproved only when χ^2 exceeds the 5 percent level, the treatment means may be tested by regarding them as approximately normally distributed with variance

$1/5 \times 0.25 = 0.05$. It is, however, more prudent to use the actual error mean square as an estimate of the experimental error variance, performing the usual tests associated with the analysis of variance. This may be justified on the grounds that the calculations have produced a set of plot values a' of equal weight. On this basis the standard error of a treatment mean is $\sqrt{0.3831/5} = 0.2768$. Treatment K reduced the number of wireworms significantly below all other treatments, but there is no indication of any difference between the other treatments. The treatment means may be reconverted to the original units by squaring.

4. Experimental variation of the binomial type. In this case the yields are obtained by examining a constant number n units per plot and noting those which possess a certain attribute (e.g., plants which are diseased). Experimental variation is presumed to arise solely from the binomial variation of the observed fraction p possessing the attribute about the expected fraction P, which is specified in terms of unknown parameters representing the treatment and soil effects.

If r_i is the number possessing the attribute on a typical plot, so that $p_i = r_i/n$ the likelihood function takes the form

$$\prod_i \frac{n!}{r_i!(n-r_i)!} P_i^{r_i} Q_i^{n-r_i}.$$

Hence the terms in the logarithm which involve the unknown parameters are given by

$$L = \sum_i \{r_i \log P_i + (n - r_i) \log Q_i\}. \tag{15}$$

The equation of estimation for a typical constant θ is

$$\Sigma \frac{n}{P_i Q_i} (p_i - P_i) \frac{\partial P_i}{\partial \theta} = 0 \tag{16}$$

where the summation is over all plots whose expectations involve θ.

As in the Poisson case, an exact solution is laborious because of the weights $\frac{n}{P_i Q_i} \cdot \frac{\partial P_i}{\partial \theta}$. The unequal weighting may be removed by transforming to the variate $\alpha_i = \sin^{-1} \sqrt{P_i}$, and assuming that the prediction formula is linear in the transformed scale. For a Latin square the prediction formula is assumed to be

$$\alpha_i = G + T_t + R_r + C_c \tag{17}$$

where the ith plot receives treatment t and lies in the rth row and cth column. Further

$$\sum_t T_t = \sum_r R_r = \sum_c C_c = 0. \tag{18}$$

Since $P_i = \sin^2 \alpha_i$, $\frac{dP_i}{d\alpha_i} = 2\sqrt{P_iQ_i}$. A set of variates a_i' is defined so that on each plot

$$p_i - P_i = (a_i' - \alpha_i)\frac{dP_i}{d\alpha_i} = 2\sqrt{P_iQ_i}\,(a_i' - \alpha_i). \tag{19}$$

With these substitutions, the equation of estimation for T_t, for instance, becomes

$$\sum_{T_t} 4n(a_i' - \alpha_i) + \lambda = 0 \tag{20}$$

where, as before, λ is an undetermined multiplier. The remainder of the solution proceeds exactly as in the Poisson case, T_t being found to be the difference between the mean value of a_i' over all plots receiving this treatment and the general mean of a_i'. A χ^2 test may be made with $\sum_i 4n(a_i' - \alpha_i)^2$.

From (19)

$$a_i' = \alpha_i + \frac{1}{2\sqrt{P_iQ_i}}(p_i - P_i) = \alpha_i + \frac{1}{2\sqrt{P_iQ_i}}(Q_i - q_i) \tag{21}$$

$$= \alpha_i + \tfrac{1}{2}\cot \alpha_i - q_i \operatorname{cosec}(2\alpha_i) \tag{22}$$

where q_i is the observed fraction which does not possess the attribute. The calculation of approximations to a_i' thus involves finding a predicted value α_i from the treatment and block (or row and column) means, and using equation (22). Tables [10] of the values of $\sin^{-1}\sqrt{P_i}$, $\alpha_i + \frac{1}{2}\cot \alpha_i$, and $\operatorname{cosec}(2\alpha_i)$ have been prepared to facilitate the computations. It should be noted that these tables are in degrees, whereas the above equations assume that α_i is measured in radians. In degrees, equation (20) above becomes

$$\sum_{T_t} \frac{\pi^2 n}{8100}(a_i' - \alpha_i) = 0 \tag{23}$$

while

$$a_i' = \alpha_i + \frac{180}{\pi}\{\tfrac{1}{2}\cot \alpha_i - q_i \operatorname{cosec}(2\alpha_i)\}. \tag{24}$$

As in the Poisson case, the appropriateness of the linearly additive law in equivalent angles depends on the way in which treatment and soil effects operate. As Bliss has shown [11], the effect of the transformation is to flatten out the cumulative normal frequency distribution, extending the range over which it can be approximated by a straight line.

5. Numerical example of the angular transformation. The data were selected from a randomized blocks experiment by Carruth [12] on the control by mechanical and insecticidal methods of damage due to corn ear worm larvae.

The control and the six types of mechanical protection were chosen for analysis, the "yields" being the percentages of ears unfit for sale. The numbers of ears varied somewhat from plot to plot, the average being 36.5, but the variations were fairly small and appeared to be random. It was considered that variations in the weight ($4n$) could be ignored in solving the equations of estimation.

TABLE III

Percentages of unfit ears of corn

Treatments	Blocks						Means
	I	II	III	IV	V	VI	
	42.4[1]	34.3	24.1	39.5	55.5	49.1	
1	40.6[2]	35.8	29.4	38.9	48.2	44.5	39.57[2]
	40.7[3]	36.0	29.4	38.9	48.6	44.6	39.70[3]
	23.5	15.1	11.8	9.4	31.7	15.9	
2	29.0	22.9	20.1	17.9	34.3	23.5	24.62
	29.1	23.1	20.3	18.2	34.3	23.5	24.75
	33.3	33.3	5.0	26.3	30.2	28.6	
3	35.2	35.2	12.9	30.9	33.3	32.3	29.97
	35.5	35.3	14.5	31.0	33.4	32.4	30.35
	11.4	13.5	2.5	16.6	39.4	11.1	
4	19.7	21.6	9.1	24.0	38.9	19.5	22.13
	19.8	21.7	10.0	24.4	39.9	19.6	22.57
	14.3	29.0	10.8	21.9	30.8	15.0	
5	22.2	32.6	19.2	27.9	33.7	22.8	26.40
	22.6	32.7	19.2	28.0	33.7	22.9	26.52
	8.5	21.9	6.2	16.0	13.5	15.4	
6	17.0	27.9	14.4	23.6	21.6	23.1	21.27
	17.4	28.2	14.5	24.0	22.1	23.2	21.57
	16.6	19.3	16.6	2.1	11.1	11.1	
7	24.0	26.1	24.0	8.3	19.5	19.5	20.23
	24.3	26.2	28.8	10.9	20.1	19.5	21.63
Means	26.81[2]	28.87	18.44	24.50	32.79	26.46	26.31

[1] Percentage. [2] Equivalent angle. [3] Second approximation.

The percentages of unfit ears, the equivalent angles and the second approximations to a' are shown in descending order in Table III. The percentages on

individual plots vary from 2.1 to 55.5. The second approximations were calculated from the block and treatment means of the angles. For the control plot (treatment 1) in block I, for example, the expected value is

$$39.57 + 26.81 - 26.31 = 40.07.$$

Since Fisher and Yates's tables of $\alpha + \frac{1}{2} \cot \alpha$ and cosec (2α) are given for values of α from 45° to 90°, we take the complement of the expected value, which is 49.93. Interpolating mentally from the table, we find

$$\alpha + \tfrac{1}{2} \cot \alpha = 74.0, \text{ cosec } (2\alpha) = 58.3.$$

Thus the second approximation to the complement of the angle is

$$74.0 - 0.424 \times 58.3 = 49.3.$$

Hence the second approximation to α' is 40.7, which agrees very closely with the equivalent angle.

On the majority of the plots, the second approximation differs by only a trivial amount from the equivalent angle. The plots with the three lowest percentages (2.1, 2.5, and 5.0) have increased somewhat more, and also one or two other plots where the angles deviated considerably from the expected values. A third set of approximations was not considered necessary.

The analysis of variance of the second approximations is given in Table IV.

TABLE IV

	Degrees of freedom	Sum of squares	Mean squares
Blocks	5	709.79	
Treatments	6	1,531.56	255.26
Error	30	982.67	32.76

Taking n as 36.5, the expected value of the error mean square is 820.7/36.5 = 22.48. Thus $\chi^2 = 982.67/22.48 = 43.71$, with 30 degrees of freedom, which is almost exactly at the 5 percent level. This, together with the appreciable amount of the variance removed by blocks, indicates that the experimental error probably contains some element other than binomial variation. As in the preceding case, it would be wise to make the usual analysis of variance tests with the actual error mean square.

6. Discussion. It must be emphasized that the solutions given above apply to the case where the whole of the experimental error variation is of the Poisson or binomial type. The methods are therefore likely to be useful in practice only where the experimental conditions have been carefully controlled, or where the data are derived from such small numbers that the Poisson or binomial variation is much larger than any extraneous variation. The χ^2 test is helpful in deciding

whether this assumption is justified. Further, the examples worked above indicate that the transformed values form very good approximations on most plots. It will often be sufficient to adjust only those plots which give zero or very small values in the Poisson case, or zero or 100 percent values in the binomial case. In this connection the method of adjustment given above may perhaps be considered as an improvement on the empirical rule given by Bartlett [13] of counting n out of n as $(n - 1/4)$ out of n.

Where extraneous variation becomes important, as is probably the normal case with data derived from field experiments, there seem to be no theoretical grounds for using the adjusted values. If we were prepared to describe accurately the nature of the variation other than that of the Poisson or binomial type, a new set of maximum likelihood equations could be developed. These would, however, lead to a different type of adjustment.

The justification for the use of transformations has no direct relation to the Poisson or binomial laws in this case, or in cases where percentages are derived from the ratios of two *weights* or volumes, as in chemical analyses, or from an arbitrary observational scoring With percentages, for example, it may be said, without describing the experimental variation in detail, that the variance must vanish at zero and 100 percent and is likely to be greatest in the middle. The formula $V = \lambda PQ$ is at least a first approximation to this situation. The angular transformation will approximately equalize a distribution of variances of this type, provided that λ is sufficiently small. We have, of course, returned to an "approximate" type of argument. It follows that the original data should be scrutinized carefully before deciding that a transformation is necessary and that any presumed opinions about the nature of the experimental variation should be verified as far as possible.

7. Summary. This paper discusses the theoretical basis for the use of the square root and inverse sine transformations in analyzing data whose experimental errors follow the Poisson and binomial frequency laws respectively.

The maximum likelihood equations of estimation are developed for each case, but are in general too complicated for frequent use. If, however, the expected yield of any plot is assumed to be an additive function of the treatment and soil effects in the transformed scale, a transformation can be found so that the equations of estimation assume the simple "normal theory" form. The transforms are closely related to the square roots and inverse sines respectively.

The nature of the assumed formula for the expected values is briefly discussed, and a χ^2 test is developed for the combined hypotheses that the prediction formula is satisfactory and that the experimental errors follow the assumed law.

Numerical examples are worked for both types of transformation. These indicate that even for data derived from small numbers, the square roots or inverse sines are good estimates of the correct transforms on almost all plots, except those which give zero yields in the Poisson case, or percentages near zero or 100 in the binomial case.

In practice, these new methods are not recommended to supplant the simple transformations for general use, because it can seldom be assumed that the whole of the experimental error variation follows the Poisson or binomial laws. The more exact analysis may, however, be useful (*i*) for cases in which the plot yields are very small integers or the ratios of very small integers (*ii*) in showing how to give proper weight to an occasional zero plot yield.

REFERENCES

[1] W. G. COCHRAN, "Some difficulties in the statistical analysis of replicated experiments," *Empire J. Expt. Agric.*, Vol. 6 (1938), pp. 157–75.

[2] M. S. BARTLETT, "The square root transformation in the analysis of variance," *J. Roy. Stat. Soc. Suppl.*, Vol. 3 (1936), pp. 68–78.

[3] C. I. BLISS, "The transformation of percentages for use in the analysis of variance," *Ohio J. Sci.*, Vol. 38 (1938), pp. 9–12.

[4] A. CLARK AND W. H. LEONARD, "The analysis of variance with special reference to data expressed as percentages," *J. Amer. Soc. Agron.*, Vol. 31 (1939), pp. 55–56.

[5] W. G. COCHRAN, "The χ^2 distribution for the Binomial and Poisson series, with small expectations," *Ann. Eugen.*, Vol. 7 (1936), pp. 207–17.

[6] P. V. SUKHATME, "On the distribution of χ^2 in samples of the Poisson series," *J. Roy. Stat. Soc. Suppl.*, Vol. 5 (1938), pp. 75–9.

[7] C. I. BLISS, "The determination of the dosage-mortality curve from small numbers," *Quart. J. Pharmacy and Pharmacology*, Vol. 11 (1938), pp. 192–216.

[8] A. W. JONES, "Practical field methods of sampling soil for wireworms," *J. Agric. Res.*, Vol. 54 (1937), pp. 123–34.

[9] W. G. COCHRAN, "The information supplied by the sampling results," *Ann. App. Biol.*, Vol. 25 (1938), pp. 383–9.

[10] R. A. FISHER AND F. YATES, *Statistical tables for agricultural, biological and medical research*, Edinburgh, Oliver and Boyd, 1938.

[11] C. I. BLISS, "The analysis of field experimental data expressed in percentages," *Plant Protection* (Leningrad), 1937, pp. 67–77.

[12] L. A. CARRUTH, "Experiments for the control of larvae of *Heliothis Obsoleta Fabr*," *J. Econ. Ent.*, Vol. 29 (1936), pp. 205–9.

[13] M. S. BARTLETT, "Some examples of statistical methods of research in agriculture and applied biology," *J. Roy. Stat. Soc. Suppl.*, Vol. 4 (1937), p. 168, footnote.

IOWA STATE COLLEGE,
AMES, IOWA