THE TRANSFORMATION OF PERCENTAGES FOR USE IN THE ANALYSIS OF VARIANCE

C. I. BLISS,
Institute of Plant Protection,
Leningrad, U. S. S. R.

The unit used most frequently in expressing the results of field experiments in economic entomology is the percentage, such as the percentage mortality, the percentage infestation, etc. Usually these experiments have not been planned so as to include an objective estimate of the experimental error but recently several writers[1] have attempted to correct this deficiency by means of the analysis of variance. Although the analysis of variance probably serves this purpose better than any other method, it was not developed originally for use with percentages and it is desirable to examine this application somewhat more closely.

Two essential features of the analysis of variance are (1) that the plots containing different treatments are exposed equally and at random to the chance of experimental error, and (2) that all contributions to the net experimental error are pooled to give a single estimate of its magnitude, with which the variation due to treatment can be compared. Presumably the portion of this error coming from each plot is independent of the treatment to which it has been exposed, so that all plots contribute equally. However, when the experimental results are in terms of percentages, the error is a function not only of the number of individuals upon which the percentage is based—which often can be equalized experimentally—but also of the theoretical percentage which is sampled by the observed value. If, in fact, all treatments were to produce the same percentage effect within the limits of the sampling error, so that the theoretical percentage could be taken as constant, then the pooled estimate of error would be a valid one and could not lead to incorrect conclusions. But more often the treatments will not be of equal effectiveness and in such cases the results on plots that are given some treatments will be estimated within narrower limits than the results on plots that are given other treatments.

[1] L. L. Huber and J. P. Sleesman, J. Econ. Entom., 28, 70, 1935. T. R. Hansberry and C. H. Richardson, Ia. State Col. J. Science, 10,27, 1935.

Under these circumstances a pooled estimate of error, especially in the study of interactions, may not be a reliable measuring stick. The discrepancy is further increased if a significant amount of field heterogeneity has been eliminated from the estimate of the experimental error by a randomized block or Latin square arrangement.

The information, I_p, in an observed percentage (or proportion) is by definition the reciprocal of its variance and for large samples is given by the equation

$$I_p = \frac{n}{pq},$$

where p is the theoretical or expected proportion of one type of individual, such as of dead insects after a poison spray, $q = 1 - p$ or the theoretical proportion of the alternative type, as of insects surviving the spray, and n is the number of individuals counted in determining a given percentage. The dependence of I_p upon this theoretical proportion is shown in Fig. 1, in which I_p is taken as unity when $p = 0.5$. The information contained in any observation is a minimum at $p = 0.5$ and increases rapidly as the proportion falls below 0.1 or rises above 0.9. Moreover, the theoretical proportion is not known *a priori* but is itself the object of estimation.

The most convenient way of eliminating this variability would be to transform the observed percentages to a unit that is not dependent upon the theoretical proportion but, whatever its value, contains an equal amount of information. Such a procedure would be analogous to the conversion of the correlation coefficient to the statistic z when testing significance or when combining data, as described in Section 35 of "Statistical Methods for Research Workers," by R. A. Fisher. The transformation of percentages to probits,[2] which were introduced for another purpose, would not meet our requirements, since the information on the probit scale is also dependent upon the expected proportion, although it has a maximum instead of a minimum at $p = 0.5$ (Fig. 1).

Prof. R. A. Fisher has written me that the problem can be resolved by transforming each percentage to an angle θ such that $p = \sin^2\theta$. As the proportion p varies from 0 to 1 or the observed percentage from 0 to 100, θ will change from 0° to 90°.

[2] C. I. Bliss, Ann. Appl. Biol. **22,** 134. 1935.

Then

$$q = \cos^2 \theta,$$

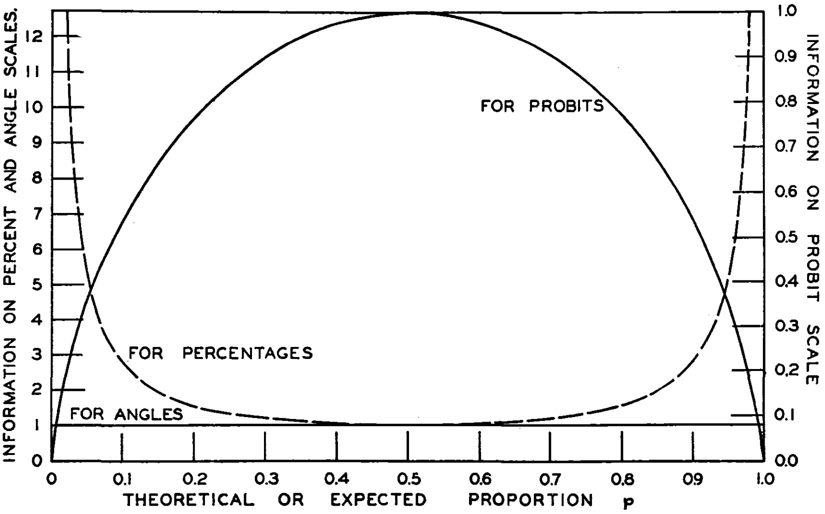
$$I_p = \frac{n}{pq} = \frac{n}{\sin^2 \theta \cos^2 \theta},$$

and

$$I_\theta = I_p\left(\frac{dp}{d\theta}\right)^2 = \frac{n}{\sin^2 \theta \cos^2 \theta} \,.\, 4 \sin^2 \theta \cos^2 \theta = 4n,$$

since

$$\frac{dp}{d\theta} = 2 \sin \theta \cos \theta.$$



The value for I_θ is exact and is independent of p or θ, as indicated by the horizontal line in Fig. 1. For large samples the distribution tends to normality and consequently has the limiting variance formula $V(\theta) = \frac{1}{4}n$. In most field experiments the percentages are based upon relatively large numbers, of 100 or more individuals, so that usually the transformation would accomplish its purpose. For small samples the distribution is not normal and the effect that this may have upon the variance of the angle when n = 10 (which may be taken as a minimum) has been discussed in a recent paper by M. S. Bartlett.[3] His study shows that at this lower limit the variance is still a function of p, but not more so than is the original percentage.

[3] M. S. Bartlett, J. Roy. Stat. Soc. Supplement **3**, 68. 1936.

Although the gain would be less, the use of this transformation for samples of only moderate size should not introduce any new errors in the subsequent analysis of variance.

The usefulness of the transformation from percentages to equivalent angles depends upon the availability of tables by which the one can be converted directly into the other. It is essential for later computation in the analysis of variance that the fractions of these equivalent angles be expressed in a decimal system rather than in minutes and seconds. Such a table has been computed. It will be published elsewhere in full[4] but is given here in abbreviated form. With this aid it should be possible to apply the analysis of variance without error to such field experimental data as can be expressed legitimately in percentages, even when these cover a wide range of values.

TABLE I

Angles of Equal Information are Given in the Body of the Table Corresponding to Observed Percentages Along the Left Margin and Top

	0	1	2	3	4	5	6	7	8	9
0	0	5.7	8.1	10.0	11.5	12.9	14.2	15.3	16.4	17.5
10	18.4	19.4	20.3	21.1	22.0	22.8	23.6	24.4	25.1	25.8
20	26.6	27.3	28.0	28.7	29.3	30.0	30.7	31.3	31.9	32.6
30	33.2	33.8	34.4	35.1	35.7	36.3	36.9	37.5	38.1	38.6
40	39.2	39.8	40.4	41.0	41.6	42.1	42.7	43.3	43.9	44.4
50	45.0	45.6	46.1	46.7	47.3	47.9	48.4	49.0	49.6	50.2
60	50.8	51.4	51.9	52.5	53.1	53.7	54.3	54.9	55.6	56.2
70	56.8	57.4	58.1	58.7	59.3	60.0	60.7	61.3	62.0	62.7
80	63.4	64.2	64.9	65.6	66.4	67.2	68.0	68.9	69.7	70.6
90	71.6	72.5	73.6	74.7	75.8	77.1	78.5	80.0	81.9	84.3
100	90.0									

[4] C. I. Bliss, Plant Protection, No. 12. 1937. Leningrad.