# 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 $i$th plot, which receives the $t$th treatment and occurs in the $r$th row and the $c$th 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

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