ANALYSIS OF VARIANCE CONSIDERED AS AN APPLICATION OF SIMPLE ERROR THEORY

By Walter A. Hendricks

The need for an elementary presentation of the methods of analysis of variance has been recognized by many investigators in various fields of research. A recent monograph by Snedecor (1934) is undoubtedly the most comprehensive attempt to satisfy this need which has appeared in the literature relating to the subject. Snedecor's treatment of the subject consists largely of the presentation of a number of standard types of problems to which the methods of analysis of variance are applicable, directions for performing the necessary computations, and a discussion of the conclusions which may be drawn from the data on the basis of the analysis.

In the opinion of the author of this paper, an elementary presentation of some of the theoretical considerations upon which the methods of analysis of variance are based would also be of some value. The methods of analysis of variance, as given by Fisher (1932), are presented as a natural consequence of intraclass correlation theory. However, the essential concepts may be presented in a more comprehensible form by the use of simple error theory.

It seems appropriate to begin such a presentation with a definition of variance. If we have an infinite number of measurements of the same quantity, the variance of a single measurement is defined as the arithmetic mean of the squares of the errors of those measurements. In actual practice, an infinite number of measurements can never be obtained. We have instead a sample of n measurements, $x_1, x_2, \cdots x_n$, from which the variance of a single measurement may be estimated. By referring to any text on the method of least squares, it may be verified that the best estimate, S^2, of the variance of a single measurement which can be obtained from a sample of n measurements is given by the equation:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

in which m represents the arithmetic mean of the n measurements. The quantity, $n - 1$, in the terminology of analysis of variance, is designated as the number of degrees of freedom available for estimating S^2.

It is often necessary to estimate S^2 from a number of different samples of measurements. In such cases, the best estimate of S^2 is obtained by calculating the weighted mean of the variances estimated from the individual samples, each variance being weighted by the number of degrees of freedom which were avail-

able for its estimation. The number of degrees of freedom upon which such an estimate of S^2 is based is given by the sum of these weights. Such an estimate of the variance of a single measurement is often designated as the variance "within samples."

In one of the simpler applications of analysis of variance, a number of samples of measurements are available, and the investigator is required to determine whether the magnitude of the quantity measured varied from sample to sample or whether all of the measurements may be regarded as having been made upon a quantity of the same magnitude.

An estimate, S^2, of the variance within samples may be obtained. Since S^2 is an estimate of the variance of a single measurement, the variance, S_i^2, of the arithmetic mean, m_i, of the measurements in any one sample is given by the equation:

$$S_i^2 = \frac{S^2}{n_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

in which n_i represents the number of measurements in the sample. Let there be r samples. Then another estimate, $S_i'^2$, of the variance of the mean, m_i, may be obtained from the observed distribution of the means, $m_1, m_2, \cdots m_r$, by the use of the formula for calculating the variance of a weighted observation as given in texts on the method of least squares:

$$S_i'^2 = \frac{1}{n_i(r-1)}[n_1(m_1 - m)^2 + n_2(m_2 - m)^2 + \cdots + n_r(m_r - m)^2] \ldots . (3)$$

in which:

$$m = \frac{n_1 m_1 + n_2 m_2 + \cdots + n_r m_r}{n_1 + n_2 + \cdots + n_r} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Equations (2) and (3) yield two estimates of the variance of the mean, m_i. It is apparent that these two estimates will be equal, within the limits of sampling fluctuations, if all of the measurements in the r samples were made upon a quantity of the same magnitude. If the magnitude of the quantity measured varied from sample to sample, $S_i'^2$ will be greater than S_i^2. However, in actual practice, the two estimates of the variance of a particular mean are not compared directly. An equivalent comparison is made between two estimates of the variance of a single measurement. The first of these is nothing more than the variance within samples discussed earlier in this paper. The second estimate, which may be designated by S'^2, is the value which would have to be substituted for S^2 in equation (2) in order to make S_i^2 equal to the value given for $S_i'^2$ by equation (3). It is quite apparent that S'^2 may be found by the use of the equation:

$$S'^2 = \frac{1}{r-1}[n_1(m_1 - m)^2 + n_2(m_2 - m)^2 + \cdots + n_r(m_r - m)^2] . \ldots . (5)$$

S'^2 is often designated as the variance "between samples." A comparison of S'^2 with S^2 is obviously equivalent to a comparison of $S_i'^2$ with S_i^2.

If S'^2 is greater than S^2, a statistic, z, may be calculated:

$$z = \frac{1}{2} \log_e \frac{S'^2}{S^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

This statistic serves as a useful comparison between S'^2 and S^2 since its sampling distribution is known if all of the measurements comprising the data under investigation were made upon a quantity of the same magnitude. The distribution of z, under these conditions, is given by an equation of the form:

$$df = \frac{ke^{n_1 z}}{(n_1 e^{2z} + n_2)^{\frac{1}{2}(n_1+n_2)}} dz \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

in which n_1 represents the number of degrees of freedom available for estimating S'^2, and n_2 represents the number of degrees of freedom available for estimating S^2. It is apparent from equation (5) that $r - 1$ degrees of freedom are available for the estimation of S'^2 in the particular problem under discussion.

When any estimate of the variance of a single measurement is multiplied by the number of degrees of freedom available for making that estimate, the resulting product is known as a "sum of squares." The additive property of the sums of squares and the degrees of freedom contributes much to the elegance of the scheme of analysis just presented and is of considerable practical importance in problems of a type to be discussed later in this paper. In the case of the problem discussed above, the additive property of the sums of squares provides that the sum of the "sum of squares between samples" and the "sum of squares within samples" is equal to the sum of the squares of the deviations of all of the measurements from their arithmetic mean. The additive property of the degrees of freedom provides that the sum of the "degrees of freedom between samples" and the "degrees of freedom within samples" is equal to the "total degrees of freedom" which is nothing more than the total number of measurements diminished by unity.

The methods of analysis presented above may be applied to any study of the effects of a number of experimental treatments of the same kind upon the magnitude of a measurable quantity. If experimental treatments of more than one kind are imposed simultaneously, the effects of each may be studied by modifications of those methods. The discussion of those modifications, about to be presented in this paper, is limited to data which may be classified in an "$r \times s$" table, i.e., to studies of the effects of only two kinds of experimental treatments. More complex problems may be treated by simple extensions of the methods presented.

Consider an "$r \times s$" table composed of rs cells, each of which contains a number of measurements of some quantity. The magnitude of the quantity measured may vary from cell to cell, but the essential conditions under which the measurements were made must be the same for all cells. It is also under-

stood that no cell may be empty. Table 1 is an example of such a table. The individual measurements have not been represented. Only the number of measurements, n_{ij}, in each cell and the arithmetic mean, m_{ij}, of those measurements have been indicated. The arguments, a_i, represent r experimental treatments of one kind, and the arguments, b_j, represent s experimental treatments of another kind. The problem to be solved is to ascertain whether or not the differences among the experimental treatments of each kind had any effect on the magnitude of the quantity measured.

TABLE 1

Example of an "$r \times s$" Table Showing Only the Number of Measurements in Each Cell and the Arithmetic Mean of Those Measurements

	b_1	b_2	b_3	b_4		b_s
a_1	m_{11} n_{11}	m_{12} n_{12}	m_{13} n_{13}	m_{14} n_{14}		m_{1s} n_{1s}
a_2	m_{21} n_{21}	m_{22} n_{22}	m_{23} n_{23}	m_{24} n_{24}		m_{2s} n_{2s}
a_3	m_{31} n_{31}	m_{32} n_{32}	m_{33} n_{33}	m_{34} n_{34}		m_{3s} n_{3s}
a_r	m_{r1} n_{r1}	m_{r2} n_{r2}	m_{r3} n_{r3}	m_{r4} n_{r4}		m_{rs} n_{rs}

If each cell contains the same number of measurements, the effects of the experimental treatments indicated by the arguments, a_i, may be studied by comparing the variance "between rows" with the variance "within cells." The variance between rows may be calculated by regarding the r rows as r samples of measurements and applying an equation of the same form as equation (5). The variance within cells may be obtained by calculating the variance of a single measurement from the data in each cell separately and taking the mean of the resulting values. The effects of the experimental treatments indicated by the arguments, b_j, may be studied by comparing the variance "between columns" with the variance "within cells."

If the degrees of freedom between rows, between columns, and within cells are added, the sum will be less than the total number of degrees of freedom in the table. If the corresponding sums of squares are added, the sum is likely to be less than the total sum of squares. The differences are due to what is customarily designated as "interaction between rows and columns." The

more descriptive term, "differential response," is sometimes used to designate the same factor. The nature of this factor may be investigated by considering the effects of the experimental treatments, b_j, in each row of Table 1.

The data in each cell of Table 1 may be regarded as a sample of measurements. Therefore, the data in any row may be regarded as a set of s samples of measurements. By applying an equation of the same form as equation (5) to the data in any row, an estimate of the variance of a single measurement is obtained from the observed distribution of the means of the cells in that row. By calculating the arithmetic mean of the estimates for the r rows, an estimate of the variance of a single measurement is obtained from $r(s - 1)$ degrees of freedom. This estimate may be designated as the variance "between cells in the same row."

The variance between cells in the same row measures the average effect of differences among the experimental treatments, b_j, in individual rows. The variance between columns, which was discussed earlier in this paper, is calculated from $s - 1$ degrees of freedom and measures the effect of differences among the treatments, b_j, on the assumption that the effect of any one treatment upon the magnitude of the quantity measured was constant for every row. The number of degrees of freedom assignable to differential response of the various rows to the treatments, b_j, is $r(s - 1) - (s - 1)$ or $(r - 1)(s - 1)$. The sum of squares due to differential response is given by the difference between the sum of squares between cells in the same row and the sum of squares between columns. These relations follow from the additive property of degrees of freedom and sums of squares.

It may be observed that precisely the same results would be obtained by considering the effects of the treatments, a_i, in the various columns of Table 1. The degrees of freedom and sum of squares due to differential response of the various columns to the treatments, a_i, would be exactly equal to the corresponding values obtained for the differential response of the various rows to the treatments, b_j.

Up to this point the discussion has been concerned only with the special case in which each cell of Table 1 contains the same number of measurements. As a matter of fact, the methods given for the analysis of such data will yield correct results when applied to any "$r \times s$" table in which the numbers of measurements in the cells in every row are proportional to the corresponding marginal totals for the columns, and the numbers of measurements in the cells in every column are proportional to the corresponding marginal totals for the rows.

When the numbers of measurements in the various cells do not satisfy the above condition of proportionality, the distributions of the means of the rows and columns may be distorted, and, consequently, the methods of analysis described above may yield incorrect results. Efficient methods of analyzing such data have been presented by Yates (1933). A comprehensive discussion of these methods is considerably beyond the scope of this paper. One method,

described very briefly by Yates (1933) and designated as the "method of weighted squares of means," appealed to the author as being particularly valuable for practical work. No detailed discussion of the method seems to be available in the literature. Therefore, the following presentation may be of some interest.

Consider the experimental treatments represented by the arguments, a_i, in Table 1. It is necessary to find an average value for the magnitude of the quantity measured for each row of Table 1. However, this average must be of such a type that its value will not be distorted by the unequal numbers of measurements in the various cells. The unweighted arithmetic mean of the means of the cells in the row seems to be the logical average to use since, within the limits of sampling fluctuations, the value of this average will be identical with the value which would have been obtained if each cell had contained the same number of measurements. The averages for the r rows are:

$$\begin{aligned} m_a &= \frac{1}{s}(m_{11} + m_{12} + \cdots + m_{1s}) \\ m_{a_2} &= \frac{1}{s}(m_{21} + m_{22} + \cdots + m_{2s}) \\ &\vdots \\ m_{a_r} &= \frac{1}{s}(m_{r1} + m_{r2} + \cdots + m_{rs}). \end{aligned} \qquad \text{.....(8)}$$

By the law of propagation of error, the variance of any one of these unweighted means is given by the equation:

$$S^2_{a_i} = \frac{1}{s^2}(S^2_{i1} + S^2_{i2} + \cdots + S^2_{is}) \qquad \text{.....(9)}$$

in which $S^2_{a_i}$ is the variance of m_{a_i}, and $S^2_{i1}, S^2_{i2}, \cdots, S^2_{is}$ are the variances of $m_{i1}, m_{i2}, \cdots, m_{is}$, respectively. If S^2 represents the variance of a single measurement, equation (9) may be written in the form:

$$S^2_{a_i} = \left(\frac{1}{n_{i1}} + \frac{1}{n_{i2}} + \cdots + \frac{1}{n_{is}}\right)\frac{S^2}{s^2} \qquad \text{.....(10)}$$

The value of S^2 may be estimated from the individual measurements in the various cells. S^2 is nothing more than the variance within cells, as customarily calculated, and may be estimated from the $N - rs$ degrees of freedom within cells, in which N represents the total number of measurements in Table 1.

The variance of a single measurement may also be estimated from the observed distribution of the means of the type, m_{a_i}. These means are not of equal weight. Therefore, in order to find the variance of any one of them, it is first necessary to calculate the weighted mean of the r individual means. Since the weight of an arithmetic mean is inversely proportional to its variance, it is evident from

an inspection of equation (10) that the weight, p_{a_i}, of a mean, m_{a_i}, may be found from the equation:

$$\frac{1}{p_{a_i}} = \frac{1}{n_{i1}} + \frac{1}{n_{i2}} + \cdots + \frac{1}{n_{is}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (11)$$

The weighted mean, m_a, may then be found:

$$m_a = \frac{p_{a_1} m_{a_1} + p_{a_2} m_{a_2} + \cdots + p_{a_r} m_{a_r}}{p_{a_1} + p_{a_2} + \cdots + p_{a_r}} \quad \ldots\ldots\ldots\ldots\ldots (12)$$

The variance $S'^2_{a_i}$, of any mean, m_{a_i}, as estimated from the observed distribution of means of this type, is given by:

$$S'^2_{a_i} = \frac{1}{p_{a_i}(r-1)} [p_{a_1}(m_{a_1} - m_a)^2 + p_{a_2}(m_{a_2} - m_a)^2 + \cdots + p_{a_r}(m_{a_r} - m_a)^2]. \quad \ldots\ldots (13)$$

By substituting $S'^2_{a_i}$ for $S^2_{a_i}$, and S^2_a for S^2, in equation (10) and solving the resulting equation for S^2_a, an estimate, S^2_a, of the variance of a single measurement is obtained from the observed distribution of means of the type, m_{a_i}. It is evident that, after making the indicated substitutions, equation (10) reduces to the form:

$$S^2_a = \frac{s^2}{r-1} [p_{a_1}(m_{a_1} - m_a)^2 + p_{a_2}(m_{a_2} - m_a)^2 + \cdots + p_{a_r}(m_{a_r} - m_a)^2] \ldots . (14)$$

It is interesting to observe that, if the numbers of measurements in the respective cells were equal, equation (14) would reduce to the formula for calculating the variance "between rows" as customarily applied in analysis of variance.

The two estimates, S^2 and S^2_a, of the variance of a single measurement may be compared in the usual manner by taking one-half of the natural logarithm of the ratio of the larger estimate to the smaller and making use of the tables of the values of "z" given by Fisher (1932). When using these tables, it is important to remember that S^2_a was estimated from $r - 1$ degrees of freedom.

The method of analysis just described may be employed to study the effects of differences among the experimental treatments indicated by the arguments, b_j, on the magnitude of the quantity measured. The unweighted means for the s columns are:

$$\begin{aligned} m_{b_1} &= \frac{1}{r}(m_{11} + m_{21} + \cdots + m_{r1}) \\ m_{b_2} &= \frac{1}{r}(m_{12} + m_{22} + \cdots + m_{r2}) \\ &\vdots \qquad\qquad \vdots \\ m_{b_s} &= \frac{1}{r}(m_{1s} + m_{2s} + \cdots + m_{rs}) \quad \ldots\ldots\ldots\ldots\ldots (15) \end{aligned}$$

The weight, p_{b_j}, of a mean of the type, m_{b_j}, may be found from the relation:

$$\frac{1}{p_{b_j}} = \frac{1}{n_{1j}} + \frac{1}{n_{2j}} + \cdots + \frac{1}{n_{rj}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (16)$$

A weighted mean, m_b, may be calculated:

$$m_b = \frac{p_{b_1}m_{b_1} + p_{b_2}m_{b_2} + \cdots + p_{b_s}m_{b_s}}{p_{b_1} + p_{b_2} + \cdots + p_{b_s}} \quad \ldots\ldots\ldots\ldots\ldots (17)$$

An estimate, S_b^2, of the variance of a single measurement may be obtained from the observed distribution of means of the type, m_{b_j}, by the use of the equation:

$$S_b^2 = \frac{r^2}{s-1}[p_{b_1}(m_{b_1} - m_b)^2 + p_{b_2}(m_{b_2} - m_b)^2 + \cdots + p_{b_s}(m_{b_s} - m_b)^2] \quad \ldots\ldots (18)$$

S_b^2 may be compared with S^2 in the usual manner.

If it is necessary to study the "interaction between rows and columns," the effects of the experimental treatments, b_j, may be studied for each individual row of Table 1. Consider the distribution of the means of the cells in a row designated by the argument, a_i. The weight of any one of these means is equal to the number of measurements in the cell. A weighted mean, m'_{a_i}, of the s means of cells in the row may be calculated:

$$m'_{a_i} = \frac{n_{i1}m_{i1} + n_{i2}m_{i2} + \cdots + n_{is}m_{is}}{n_{i1} + n_{i2} + \cdots + n_{is}} \quad \ldots\ldots\ldots\ldots\ldots (19)$$

The variance, S'^2_{ij}, of the mean, m_{ij}, for any cell in the given row, as estimated from the observed distribution of means of this type, may be obtained from the equation:

$$S'^2_{ij} = \frac{1}{n_{ij}(s-1)}[n_{i1}(m_{i1} - m'_{a_i})^2 + n_{i2}(m_{i2} - m'_{a_i})^2 + \cdots + n_{is}(m_{is} - m'_{a_i})^2] \quad \ldots\ldots (20)$$

The variance, S^2_{ij}, of the same mean, as estimated from the distribution of the individual measurements in the cell, may be obtained from the equation:

$$S^2_{ij} = \frac{S^2}{n_{ij}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (21)$$

By substituting S'^2_{ij} for S^2_{ij}, and $S^2_{a_ib}$ for S^2, in equation (21) and solving the resulting equation for $S^2_{a_ib}$, an estimate, $S^2_{a_ib}$, of the variance of a single measurement is obtained from the observed distribution of the means of the cells in the given row. After making the indicated substitutions, equation (21) reduces to the form:

$$S^2_{a_ib} = \frac{1}{s-1}[n_{i1}(m_{i1} - m'_{a_i})^2 + n_{i2}(m_{i2} - m'_{a_i})^2 + \cdots + n_{is}(m_{is} - m'_{a_i})^2] \quad \ldots\ldots (22)$$

Such an estimate, $S^2_{a_i b}$, of the variance of a single measurement may be obtained for each of the r rows in Table 1. By calculating the average, $S^2_{a b}$, of the variances of the type, $S^2_{a_i b}$, an estimate, $S^2_{a b}$, of the variance of a single measurement may be obtained from the $r(s-1)$ degrees of freedom between cells in the same row:

$$S^2_{a b} = \frac{1}{r(s-1)} \sum_{i=1}^{r} [n_{i1}(m_{i1} - m'_{a_i})^2 + n_{i2}(m_{i2} - m'_{a_i})^2 + \cdots + n_{is}(m_{is} - m'_{a_i})^2] \quad \ldots\ldots(23)$$

Equation (23) is identical with the formula for calculating the variance between cells in the same row as ordinarily applied in analysis of variance. This result is a direct consequence of the fact that the unequal numbers of measurements in the various cells had no distorting effect on the arithmetic means for individual cells.

The presence or absence of interaction may be verified by comparing $S^2_{a b}$ with S^2_b. In general, the actual variance due to interaction can not be obtained by the "weighted squares of means" method, for the various sums of squares do not possess the additive property when the analysis is made in this way. However, the comparison suggested above will yield sufficient information for most practical purposes.

For the special case in which r or s is equal to 2, the actual variance due to interaction may be calculated. Suppose $r = 2$ in Table 1. The following method, suggested by Yates (1933), yields an estimate of the variance due to interaction from a consideration of the differences, d_j, between the means of the two cells in each column:

$$\begin{aligned} d_1 &= m_{11} - m_{21} \\ d_2 &= m_{12} - m_{22} \\ &\vdots \qquad\quad \vdots \\ d_s &= m_{1s} - m_{2s} \end{aligned} \quad \ldots\ldots(24)$$

The variance, $S^2_{d_j}$, of any difference, d_j, is given by the equation:

$$S^2_{d_j} = \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}}\right) S^2 \quad \ldots\ldots(25)$$

The weight, p_j, of the difference, d_j, is given by the equation:

$$\frac{1}{p_j} = \frac{1}{n_{1j}} + \frac{1}{n_{2j}} \quad \ldots\ldots(26)$$

The variance of the difference, d_j, as estimated from the observed distribution of differences, is given by the equation:

$$S'^2_{d_j} = \frac{1}{p_j(s-1)} [p_1(d_1 - d)^2 + p_2(d_2 - d)^2 + \cdots + p_s(d_s - d)^2] \quad \ldots(27)$$

in which:

$$d = \frac{p_1 d_1 + p_2 d_2 + \cdots + p_s d_s}{p_1 + p_2 + \cdots + p_s} \quad \ldots\ldots\ldots\ldots\ldots\ldots (28)$$

By means of these relations, an estimate, S_d^2, of the variance of a single measurement may be obtained from the observed distribution of the differences of the type, d_j. This estimate represents the variance due to interaction and may be obtained from the equation:

$$S_d^2 = \frac{1}{s-1}[p_1(d_1 - d)^2 + p_2(d_2 - d)^2 + \cdots + p_s(d_s - d)^2] \ldots\ldots (29)$$

It is quite apparent that $s - 1$ degrees of freedom are available for the estimation of the variance due to interaction in this particular example.

REFERENCES

FISHER, R. A., 1932. Statistical Methods for Research Workers, 4th edition. Edinburgh and London: Oliver and Boyd.

SNEDECOR, GEORGE W., 1934. Calculation and Interpretation of Analysis of Variance and Covariance. Ames, Iowa: Collegiate Press.

YATES, F., 1933. The principles of orthogonality and confounding in replicated experiments. Jour. Agr. Sci., 23: 108–145.

BUREAU OF ANIMAL INDUSTRY,
U. S. DEPARTMENT OF AGRICULTURE,
WASHINGTON, D. C.