[2] R. A. FISHER, "Two New Properties of Mathematical Likelihood." *Proc. Royal Society of London*, Series A, vol. 144(1934), pp. 285-307.
[3] R. A. FISHER, " 'Student' ". *Annals of Eugenics*, vol. 9(1939) pp. 1-9.
[4] J. NEYMAN AND E. S. PEARSON, "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Phil. Trans. Royal Society of London*, Series A, vol. 231(1933), pp. 289-337.
[5] J. NEYMAN AND E. S. PEARSON, "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses." *Statistical Research Memoirs*, vol. 1(1936), pp. 113-137.
[6] J. NEYMAN, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Phil. Trans. Royal Society of London*, Series A, vol. 236(1937), pp. 333-380.

UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN

A NOTE ON COMPUTATION FOR ANALYSIS OF VARIANCE

BY MORRIS C. BISHOP

The method of computation for analysis of variance commonly favored is one which involves obtaining the total and total sum of squares in a single operation on a computing or card-punch machine,[1] in which case a check on the accuracy of the work requires complete recomputation. But the best tools available to the student, and sometimes to the experimenter, are a table of squares and perhaps a listing machine. In such a situation, a simple algorithm which embodies checks on the computations is urgently needed. The method here presented reduces the arithmetic to repeated application of a single procedure, with adequate checks; it reveals rather than obscures the sample variances, which may or may not be of primary importance; and it provides an intuitively logical portrayal of the step-by-step improvement of the estimate of population variance.

The data items and their squares may be merged into a single table by setting them down in staggered fashion, as shown in Table I. If only a single criterion of classification is to be used—classified into columns, say—the columns are summed down, and then these totals across (obtained as two sets of subtotals and totals on a listing machine). This yields the grand total (T) and total sum of squares $\left(\sum_{i=1}^{N}\sum_{j=1}^{k} X_{ij}^2\right)$. Summing across and down verifies the addition and provides material for two-way classification. The total sum of squares of deviations is obtained by the familiar formula

$$(1) \qquad \sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{N}\sum_{j=1}^{k} X_{ij}^2 - \frac{T^2}{Nk}$$

where Nk is the total number of observations in N rows and k columns.

[1] See George W. Snedecor, *Analysis of Variance and Covariance*, and Paul R. Rider, *Modern Statistical Methods*.

TABLE I

	1	2	....	j	....	k	Row totals	Sum of squares minus $T_{j.}/k$	Sum of squared deviations within rows
1	X_{11} X_{11}^2	X_{12} X_{12}^2	....	X_{1j} X_{1j}^2	....	X_{1k} X_{1k}^2	$T_{1.}$	$\Sigma_j X_{1j}^2 - \frac{T_{1.}^2}{k}$	$\Sigma_j(X_{1j} - \overline{X}_{1.})^2$
2	X_{21} X_{21}^2	X_{22} X_{22}^2	....	X_{2j} X_{2j}^2	....	X_{2k} X_{2k}^2	$T_{2.}$	$\Sigma_j X_{2j}^2 - \frac{T_{2.}^2}{k}$	$\Sigma_j(X_{2j} - \overline{X}_{2.})^2$
.....	.....	.....	....	.....	....	.....	....	.....	.....
i	X_{i1} X_{i1}^2	X_{i2} X_{i2}^2	....	X_{ij} X_{ij}^2	....	X_{ik} X_{ik}^2	$T_{i.}$	$\Sigma_j X_{ij}^2 - \frac{T_{i.}^2}{k}$	$\Sigma_j(X_{ij} - \overline{X}_{i.})^2$
.....	.....	.....	....	.....	....	.....	....	.....	.....
N	X_{N1} X_{N1}^2	X_{N2} X_{N2}^2	....	X_{Nj} X_{Nj}^2	....	X_{Nk} X_{Nk}^2	$T_{N.}$	$\Sigma_j X_{Nj}^2 - \frac{T_{N.}^2}{k}$	$\Sigma_j(X_{Nj} - \overline{X}_{N.})^2$
Column totals	$T_{.1}$	$T_{.2}$	....	$T_{.j}$	....	$T_{.k}$	T		
Sum of squares minus $T_{.j}/N$	$\Sigma_i X_{i1}^2 - \frac{T_{.1}^2}{N}$	$\Sigma_i X_{i2}^2 - \frac{T_{.2}^2}{N}$	....	$\Sigma_i X_{ij}^2 - \frac{T_{.j}^2}{N}$	....	$\Sigma_i X_{ik}^2 - \frac{T_{.k}^2}{N}$		$\Sigma_i\Sigma_j X_{ij}^2 - \Sigma_i \frac{T_{i.}^2}{k} - \Sigma_j \frac{T_{.j}^2}{N}$	$\Sigma_i\Sigma_j(X_{ij} - \overline{X}_{i.})^2$
Sum of squared deviations within cols.	$\Sigma_i(X_{i1} - \overline{X}_{.1})^2$	$\Sigma_i(X_{i2} - \overline{X}_{.2})^2$	....	$\Sigma_i(X_{ij} - \overline{X}_{.j})^2$	....	$\Sigma_i(X_{ik} - \overline{X}_{.k})^2$		$\Sigma_i\Sigma_j(X_{ij} - \overline{X}_{.j})^2$	

To obtain the sum of squares of deviations within columns, each data column total ($T_{.j}$) is squared and divided by the number of items in the column. Then This correction $\left(\frac{T_{.j}^2}{N_j}\right)$ is subtracted from the sum of squares for the column and the differences are summed across, the formula for this summation being

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(X_{ij} - \bar{X}_{.j})^2 = \sum_{j=1}^{k}\left(\sum_{i=1}^{N_j} X_{ij}^2 - \frac{T_{.j}^2}{N_j}\right) \tag{2}$$

This procedure is unvarying, whether the classes contain the same number of observations or not. The sample variance for each column is immediately obtainable, if desired, by dividing by the appropriate degrees of freedom ($N_j - 1$). Short-cut machine methods completely obscure these sums of squares of deviations for the individual classes.

The sum of squares of deviations between classes is obtained by adding across the correction terms just used and subtracting from this total the correction term of Equation (1). That is,

$$\sum_{j=1}^{k} N_j(\bar{X}_{.j} - \bar{X})^2 = \sum_{j=1}^{k}\frac{T_{.j}^2}{N_j} - \frac{T^2}{Nk} \tag{3}$$

"Within classes" plus "Between classes" will, of course, check against the "Total".

If the classification is into rows, the appropriate formulas are:
Within rows

$$\sum_{i=1}^{N}\sum_{j=1}^{k_i}(X_{ij} - \bar{X}_{i.})^2 = \sum_{i=1}^{N}\left(\sum_{j=1}^{k_i} X_{ij}^2 - \frac{T_{i.}^2}{k_i}\right). \tag{2a}$$

Between rows

$$\sum_{i=1}^{N} k_i(\bar{X}_{i.} - \bar{X})^2 = \sum_{i=1}^{N}\frac{T_{i.}^2}{k_i} - \frac{T^2}{Nk}. \tag{3a}$$

When a single criterion of classification is involved, this procedure applies, whether or not the classes contain the same number of observations. Double classification with unequal frequencies is somewhat more difficult,[2] and we shall consider here only the case of equal numbers of observations in the rows and in the columns, respectively—a rectangular array not necessarily square.[3] For the two-way analysis, both procedures outlined above are carried out, and then

---

[2] See F. Yates, "The analysis of multiple classifications with unequal numbers in the different classes," *Journal of the American Statistical Association*, Vol. 29 (1934); and A. E. Brandt, "The analysis of variance in a 2×s table with disproportionate frequencies," *Journal of the American Statistical Association*, Vol. 28 (1933).

[3] For a simple treatment in three-way classification of unequal but proportional representation between subclasses and between classes, see G. W. Snedecor, *Statistical Methods*, p. 233-35; also an interesting example in F. C. Mills, *Statistical Methods* (1938 ed.), Appendix E.

the interaction may be obtained in any one of three ways. The usual way is to regard it as

"Total", minus "Between rows", minus "Between columns".

But it may also be considered as

"Within columns", minus "Between rows"

or

"Within rows", minus "Between columns".

Either of the latter is perhaps more meaningful than the first, because this procedure presents the interaction, or "error" sum of squares, as the logical third step in an effort to obtain the best estimate of population variance, through the successive stages of sample, or class, variances; their average; and, finally, this average freed of the effect of variability attributable to a second type of classification.

Justification for this way of regarding the interaction is easily established if we look for a moment at the familiar fundamental identity for analysis of variance:

$$\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X})^2 = N\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X})^2+\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{.j})^2$$

$$= N\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X})^2 + k\sum_{i=1}^{N}(\bar{X}_{i.}-\bar{X})^2$$

$$+\left\{\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{.j})^2 - k\sum_{i=1}^{N}(\bar{X}_{i.}-\bar{X})^2\right\}.$$

Or, similarly,

$$\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X})^2 = k\sum_{i=1}^{N}(\bar{X}_{i.}-\bar{X})^2+\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{i.})^2$$

$$= k\sum_{i=1}^{N}(\bar{X}_{i.}-\bar{X})^2 + N\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X})^2$$

$$+\left\{\sum_{i=1}^{N}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{i.})^2 - N\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X})^2\right\}.$$

In the computation notation adopted (dropping the subscripts from N and k, since class frequencies are equal), the interaction is either

$$\sum_{j=1}^{k}\left(\sum_{i=1}^{N}X_{ij}^2-\frac{T_{.j}^2}{N}\right)-\left(\sum_{i=1}^{N}\frac{T_{i.}^2}{k}-\frac{T^2}{Nk}\right)=\sum_{i=1}^{N}\sum_{j=1}^{k}X_{ij}^2-\sum_{j=1}^{k}\frac{T_{.j}^2}{N}-\sum_{i=1}^{N}\frac{T_{i.}^2}{k}+\frac{T^2}{Nk}$$

or

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{k}X_{ij}^2-\frac{T_{i.}^2}{k}\right)-\left(\sum_{j=1}^{k}\frac{T_{.j}^2}{N}-\frac{T^2}{Nk}\right)=\sum_{i=1}^{N}\sum_{j=1}^{k}X_{ij}^2-\sum_{i=1}^{N}\frac{T_{i.}^2}{k}-\sum_{j=1}^{k}\frac{T_{.j}^2}{N}+\frac{T^2}{Nk}$$

the elements of which already have been computed.

Table II demonstrates the procedure applied to a simple example involving four rows and four columns, the final computations being:
Total = 2686 − 2450.25 = 235.75
Between columns = 2461.50 − 2450.25 = 11.25
Between rows = 2463 − 2450.25 = 12.75
Interaction = 224.5 − 12.75 = 211.75, or 223 − 11.25 = 211.75

TABLE II

					Row totals		Sum of sq.dev. within rows
	8 64	10 100	12 144	16 256	46	564 − 529	35
	11 121	18 324	14 196	9 81	52	722 − 676	46
	20 400	18 324	7 49	9 81	54	854 − 729	125
	10 100	9 81	13 169	14 196	46	546 − 529	17
Column totals.	49	55	46	48	198		
	685 600.25	829 756.25	558 529	614 576		2686 − 2463 2461.5	223
Sum of sq. dev. within cols...	84.75	72.75	29	38		224.5	

The analysis of variance table is as follows:

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Total	235.75	15	
Within columns	224.50	12	18.71
Between columns	11.25	3	3.75
Within rows	223.00	12	18.58
Between rows	12.75	3	4.25
Interaction	211.75	9	23.53

Extension to more complex experimental designs involving more than two criteria of classification requires merely to observe that the interaction becomes a quantity of the order of Within A-type classes minus Between B classes minus Between C classes, as well as Total minus Between A classes minus Between B classes minus Between C classes. If, for example, the above illustration had

been designed to embody a "varieties" classification represented by the Latin square

A	C	D	B
D	B	C	A
B	D	A	C
C	A	B	D

the additional computation would be as follows:

8	64	16	256	10	100	12	144		
9	81	18	324	14	196	11	121		
7	49	20	400	9	81	18	324		
9	81	13	169	10	100	14	196		
33	275	67	1149	43	477	55	785	198	2686
	272.25		1122.25		462.25		756.25		= 2613.00
	2.75		26.75		14.75		28.75		= 73.00

Within varieties = 2.75 + 26.75 + 14.75 + 28.75 = 73.00
Between varieties = 272.25 + 1122.25 + 462.25 + 756.25 − 2450.25 = 162.75

Most complex experiments will not be concerned with the "within classes" sum of squares of deviations for all the criteria involved, but if this sum has been computed for two or more criteria, a check on the later stages of the work is had by observing the alternative ways of computing the interaction, which in this case are:

235.75	73.00	224.50	223.00
−11.25	−11.25	−12.75	−11.25
−12.75	−12.75	−162.75	−162.75
−162.75			
49.00	49.00	49.00	49.00

For the three-way classification of the Latin square the analysis of variance table is:

	Sum of squares of deviations		Degrees of freedom	Mean square deviation
Total	235.75		15	
Between varieties	162.75		3	54.25
Within varieties	73.00		12	6.08
Between columns		11.25	3	3.75
Between rows		12.75	3	4.25
Error		49.00	6	8.17

The procedure outlined, as a uniform method of computation in analysis involving a single criterion of classification, with or without equal numbers in the classes, or involving multiple criteria, has several things to recommend it. An important consideration to the worker who is not primarily a statistician is

that he or an assistant with little training can perform the mechanical work with confidence. Also, the "within classes" sum of squares is computed directly and not as a difference. The sums of squared deviations leading to the sample variances are exhibited in explicit form for inspection, and test of significance if desired. Herein frequently is found an important clue, a warning signal or a hint leading to re-examination of the sampling procedure.

As a final point, it is worthy of notice that the method facilitates use of analysis of variance as a technique of preliminary investigation. If the observed data have been obtained more or less fortuitously, and not as the result of rigid experimental design, rows or columns may be eliminated or combined with a minimum of labor, thus permitting testing of various combinations of data.

THE GEORGE WASHINGTON UNIVERSITY,
WASHINGTON, D. C.

ANNOUNCEMENT CONCERNING COMPUTATION OF MATHEMATICAL TABLES

A Project for the Computation of Mathematical Tables, sponsored by Dr. Lyman J. Briggs, Director of the National Bureau of Standards, is being conducted by the Work Projects Administration for the City of New York. The Project has been in operation since January 1, 1938, under the technical supervision of Dr. Arnold N. Lowan.

An agenda of the Project, listing tables completed, in progress and under consideration is given below:

COMPLETED TABLES

1. A table of exponentials for the following ranges, intervals and number of decimals.

Range	Interval	No. of Decimals
−2.5000 to 1.0000	0.0001	18
1.0000 to 2.5000	0.0001	15
2.500 to 5.000	0.001	15
5.00 to 10.00	0.01	12

2. A table of sines and cosines for the range from 0 to 25 radians at intervals of 10^{-3} to 8 places of decimals.
3. A table of the first 10 powers of the integers from 1 to 1,000.

TABLES IN PROGRESS

(A) Computations completed, manuscripts in process of preparation.