ON THE LEXIS THEORY AND THE ANALYSIS OF VARIANCE*

BY H. L. RIETZ

In a paper† published in 1921 involving a generalization of the Lexis theory for the classification of statistical series with regard to their dispersion into Bernoulli, Lexis, and Poisson series, Coolidge based much of his reasoning on the following fundamental dispersion theorem.

If n independent quantities $y_1, y_2, \cdots, y_n$ be given, their expected values being $a_1, a_2, \cdots, a_n$, while the expected values of their squares are $A_1, A_2, \cdots, A_n$, respectively, and if we agree to set $y = (1/n)\sum_{i=1}^{n} y_i$, $a = (1/n)\sum_{i=1}^{n} a_i$, then the expected value of the variance $(1/n)\sum_{i=1}^{n}(y_i - y)^2$ is

$$\text{(1)} \qquad \frac{1}{n}\left[\frac{n-1}{n}\sum_{i=1}^{n}(A_i - a_i^2) + \sum_{i=1}^{n}(a_i - a)^2\right].$$

In setting up criteria for the practical classification of actual statistical series, Coolidge followed the customary procedure of introducing approximations by replacing $(n-1)/n$ by 1.

By avoiding this approximation in the present paper but otherwise proceeding along the lines followed by Coolidge, we shall arrive at certain important results of R. A. Fisher in his analysis of variance. The different estimates of variance used by R. A. Fisher seem to have been obtained largely by inferences based on the number of degrees of freedom of the variates rather than upon formal mathematical proofs. In fact, the intuitional element is so prominent in certain of these inferences based on the number of degrees of freedom that mathematicians rather generally hesitate to accept the results as mathematically established, although there is much general evidence in favor of the correctness of the conclusions. For this reason, it seems of interest to show how certain of the results in question can be derived by formal developments that follow closely the reasoning of Coolidge in his generalization of the Lexis theory.

* Presented to the Society, December 28, 1931.

† This Bulletin, vol. 27 (1921), p. 439.

The Lexis theory is concerned with the examination of the inner structure of a population of items by the separation of items into subsets. As a simple mode of subdivision, suppose the set of independent items classified in some relevant manner into N sets of s items each. Then our observations may well be exhibited in columns and rows as

$$\begin{array}{lllll} x_{11}, & x_{12}, & \cdots, & x_{1s}, & \bar{x}_{1\cdot} \\ x_{21}, & x_{22}, & \cdots, & x_{2s}, & \bar{x}_{2\cdot} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{N1}, & x_{N2}, & \cdots, & x_{Ns}, & \bar{x}_{N\cdot} \\ \bar{x}_{\cdot 1}, & \bar{x}_{\cdot 2}, & \cdots, & \bar{x}_{\cdot s}, & \bar{x} \end{array} \tag{2}$$

with arithmetic means $\bar{x}_{i\cdot}$ of the ith row, $\bar{x}_{\cdot j}$ of the jth column, and $\bar{x}$ of the whole sample of Ns items.

To illustrate, take $N=17$ and $s=15$. Suppose that $N=17$ years and that $s=15$ refers to the first fifteen days of each year and that x_{ij} gives the mimimal temperature in degrees Fahrenheit at a certain place for the jth day of the ith year.

Using $E(\quad)$ for the expected value of the expression in the parenthesis, we let $E(x_{ij})=a_{ij}$, $E(x_{ij}^2)=A_{ij}$. Further, let

$$\sum_{j=1}^{s} a_{ij} = sa_i; \quad \sum_{i=1}^{N} a_i = Na; \quad \sum_{i=1}^{N} \bar{x}_{i\cdot} = N\bar{x}; \quad \sum_{j=1}^{s} \bar{x}_{\cdot j} = s\bar{x}.$$

Then by (1)

$$E\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i\cdot})^2 = \frac{(s-1)}{s}\sum_{j=1}^{s}(A_{ij}-a^{2}_{ij}) + \sum_{j=1}^{s}(a_{ij}-a_i)^2. \tag{3}$$

Summing (3) from $i=1$ to N, we have

$$E\sum_{i=1,j=1}^{N.s}(x_{ij}-\bar{x}_{i\cdot})^2 = \frac{(s-1)}{s}\sum_{i=1,j=1}^{N.s}(A_{ij}-a^{2}_{ij}) + \sum_{i=1,j=1}^{N.s}(a_{ij}-a_i)^2. \tag{4}$$

Next, we note that $E(\bar{x}_{i\cdot})=a_i$,

$$(5)\qquad \begin{aligned} \bar{x}_{i\cdot} - a_i &= \frac{1}{s}\sum_{j=1}^{s}(x_{ij} - a_{ij}), \qquad \text{and} \\ E(\bar{x}_{i\cdot} - a_i)^2 &= \frac{1}{s^2}E\left\{\sum_{j=1}^{s}(x_{ij} - a_{ij})\right\}^2 = \frac{1}{s^2}\sum_{j=1}^{s}(A_{ij} - a_{ij}^2).\end{aligned}$$ [*]

Since $E(\bar{x}_{i\cdot}) = a_i$, we note that

$$(6)\qquad E(\bar{x}_{i\cdot}^2) = E(\bar{x}_{i\cdot} - a_i)^2 + a_i^2\,.$$

By applying (1), we may write

$$(7)\ E\sum_{i=1}^{N}(\bar{x}_{i\cdot} - \bar{x})^2 = \frac{N-1}{N}\sum_{i=1}^{N}[E(x_{i\cdot}^2) - a_i^2] + \sum_{i=1}^{N}(a_i - a)^2.$$

From (6) and (7), we have

$$(8)\ E\sum_{i=1}^{N}(\bar{x}_{i\cdot} - \bar{x})^2 = \frac{N-1}{N}E\sum_{i=1}^{N}(\bar{x}_{i\cdot} - a_i)^2 + \sum_{i=1}^{N}(a_i - a)^2.$$

By substituting for $E(\bar{x}_{i\cdot} - a_i)^2$ in (8) from (5), we have

$$(9)\qquad \begin{aligned} E\sum_{i=1}^{N}(\bar{x}_{i\cdot} - \bar{x})^2 = \frac{1}{s^2}\,\frac{N-1}{N}&\sum_{i=1,j=1}^{N,s}(A_{ij} - a_{ij}^2) \\ &+ \sum_{i=1}^{N}(a_i - a)^2.\end{aligned}$$

Eliminate $\sum_{i=1,j=1}^{N,s}(A_{ij} - a_{ij}^2)$ from (4) and (9), and we obtain

$$(10)\qquad \begin{aligned} &E\sum_{i=1,j=1}^{N,s}(x_{ij} - \bar{x}_{i\cdot})^2 - \sum_{i=1,j=1}^{N,s}(a_{ij} - a_i)^2 \\ &= \frac{Ns(s-1)}{N-1}E\sum_{i=1}^{N}(\bar{x}_{i\cdot} - \bar{x})^2 - \frac{Ns(s-1)}{N-1}\sum_{i=1}^{N}(a_i - a)^2.\end{aligned}$$

Bernoulli Series. In a Bernoulli series we assume statistical homogeneity in the sense that items are so thoroughly mixed that the expected value of any statistical estimate is independent of the portion of the population from which the sample is drawn. Measurements on the same quantity will serve as an illustration, the items differing only by accidental errors. Under these conditions, $a_{ij} = a_i$, $a_i = a$. Hence, we would compare

[*] See Coolidge, *Probability*, p. 63, Theorem 8.

$$(11)\qquad \sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_{i}.)^2 \quad\text{and}\quad \frac{s(s-1)N}{N-1}\sum_{i=1}^{N}(\bar{x}_{i}.-\bar{x})^2,$$

or

$$(12)\qquad \frac{1}{N(s-1)}\sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_{i}.)^2 \quad\text{and}\quad \frac{s}{N-1}\sum_{i=1}^{N}(\bar{x}_{i}.-\bar{x})^2,$$

and expect to find them equal in a Bernoulli series except for sampling fluctuations. Those in (12) are frequently compared in the analysis of variance by the procedure of R. A. Fisher.

Lexis Series. The items within a set of s have the same expected value, but the expected values vary from one set to another. Thus, $a_{ij}=a_i$, but $a_i\neq a$. Then we expect

$$(13)\qquad \frac{1}{N(s-1)}\sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_{i}.)^2 < \frac{s}{N-1}\sum_{i=1}^{N}(\bar{x}_{i}.-\bar{x})^2.$$

The series is said to have *supernormal* dispersion.

Poisson Series. There are differences in expected values within sets of s items, but all the N sets are comparable, that is, $a_{ij}\neq a_i$; but $a_i=a$. Then we expect

$$(14)\qquad \frac{1}{N(s-1)}\sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_{i}.)^2 > \frac{s}{N-1}\sum_{i=1}^{N}(\bar{x}_{i}.-\bar{x})^2.$$

The series is said to have *subnormal* dispersion.

The expected values of the expressions in (12), (13), and (14) would be equal in a statistically homogeneous population to the expected value of $[1/(Ns-1)]\sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x})^2$ which is ordinarily regarded as the best estimate of the population variance from a sample of Ns items of such a population; that is,

$$(15)\qquad \frac{1}{N(s-1)}E\sum_{i=1,j=1}^{N.s}(x_{ij}-\bar{x}_{i}.)^2 = \frac{1}{Ns-1}E\sum_{i=1,j=1}^{N.s}(x_{ij}-\bar{x})^2.$$

To establish (15), we may start with the fact that in a statistically homogeneous system of items

$$(16)\qquad \frac{1}{s-1}\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i}.)^2$$

is our estimate of the population variance based on the use of s items drawn at random. Thus, (16) has the same expected value as the right member of (15). Furthermore, the arithmetic mean

$$\frac{1}{N(s-1)}\sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_i.)^2 \tag{17}$$

of N such independent values as (16) has the same expected value as (16). Hence (15) is established.

To summarize, we have shown with our subdivision of the items of the sample into rows and columns that the following estimates of variance have the same expected value, in a statistically homogeneous population:

$$\begin{aligned}
V &= \frac{S}{Ns-1}, &\text{where}\quad S &= \sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x})^2,\\
V_i &= \frac{S_i}{N(s-1)}, &\text{where}\quad S_i &= \sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}_i.)^2,\\
V_j &= \frac{S_j}{s(N-1)}, &\text{where}\quad S_j &= \sum_{i=1,j=1}^{N,s}(x_{ij}-\bar{x}._j)^2,\\
V_{i\bar{x}} &= \frac{sS_{i\bar{x}}}{N-1}, &\text{where}\quad S_{i\bar{x}} &= \sum_{i=1}^{N}(\bar{x}_i.-\bar{x})^2,\\
V_{j\bar{x}} &= \frac{NS_{j\bar{x}}}{s-1}, &\text{where}\quad S_{j\bar{x}} &= \sum_{j=1}^{s}(\bar{x}._j-\bar{x})^2.
\end{aligned} \tag{18}$$

We thus arrive at estimates of variance used by R. A. Fisher without making use of arguments involving the number of degrees of freedom of the items with which we are concerned. A comparison of the numerical values of $V_{i\bar{x}}$ and V_i by taking the ratio $V_{i\bar{x}}/V_i$ serves as an important step in solving the problem of testing for significant differences from row to row in (2).

Similarly, a comparison of the numerical values of $V_{j\bar{x}}$ and V_j by taking the ratio $V_{j\bar{x}}/V_j$ would be useful in testing the significance of differences from column to column in (2). In comparisons of two estimates of variance V' and V by means of the ratio V'/V, R. A. Fisher made a fundamental contribution to the theory of applied statistics by finding the distribution function of $z=\frac{1}{2}\log V'/V$ for the case of a normal parent distribution and showing it to be of such a nature that the significance of the discrepancy of z from expectation could be examined objectively and expressed in terms of odds in favor of or against a discrepancy as large as or larger than an assigned value.

The University of Iowa