A NOTE ON THE ANALYSIS OF VARIANCE[1]

By Solomon Kullback

By considering a set of independent items classified in some relevant manner into N sets of s items each, and by the use of a dispersion theorem of Prof. J. L. Coolidge,[2] Prof. H. L. Rietz[3] arrives at estimates of variance, used by Dr. R. A. Fisher, without making use of arguments involving the number of degrees of freedom of the items concerned.

By proceeding along the lines followed by Coolidge and Rietz but considering a set of independent items classified into N sets of $s_i (i = 1, 2, \cdots, N)$ items each, we shall arrive at certain other important results of R. A. Fisher[4] in his analysis of variance.

The theorem referred to above is as follows: 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

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

Suppose a set of independent items has been classified in some relevant manner into N sets of s_i $(i = 1, 2, \cdots, N)$ items each as follows:

$$\begin{array}{llllll} x_{11}, & x_{12}, & \cdots, & x_{1s_1}, & \bar{x}_1 \\ x_{21}, & x_{22}, & \cdots, & x_{2s_2}, & \bar{x}_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{N1}, & x_{N2}, & \cdots, & x_{Ns_N}, & \bar{x}_N \\ & & & & \bar{x} \end{array} \tag{2}$$

where $\bar{x}_i (i = 1, 2, \cdots, N)$ is the arithmetic mean of the i^{th} set and $\bar{x}$ the mean of the pooled sample of $s = s_1 + s_2 + \cdots + s_N$ items.

We shall assume that the set (2) is statistically homogeneous in the sense that,

---

[1] Presented to the American Mathematical Society, February 23, 1935.

[2] Bulletin Am. Math. Soc., Vol. 27 (1921) p. 439.

[3] Bulletin Am. Math. Soc., Vol. 38 (1932) pp. 731–735.

[4] Proceedings of the International Math. Congress, Toronto, 1924, Vol. 2, p. 802 ff.

using $E(\ )$ for the expected value of the expression in the parenthesis, we may let $E(x_{ij}) = a$, $E(x_{ij}^2) = A$, $(i = 1, 2, \cdots, N, j = 1, 2, \cdots, s_i)$.
Then, using (1)

$$E\left(\sum_{j=1}^{s_i} (x_{ij} - \bar{x}_i)^2\right) = (s_i - 1)(A - a^2). \tag{3}$$

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

$$E\left(\sum_{i=1, j=1}^{N, s_i} (x_{ij} - \bar{x}_i)^2\right) = (A - a^2) \sum_{i=1}^{N} (s_i - 1) = (s - N)(A - a^2). \tag{4}$$

Similarly, by using (1)

$$E\left(\sum_{i=1}^{N} s_i(\bar{x}_i - \bar{x})^2\right) = \frac{N-1}{N} \sum_{i=1}^{N} s_i [E(\bar{x}_i^2) - a^2]. \tag{5}$$

But[5]

$$E(\bar{x}_i^2) - a^2 = E(\bar{x}_i - a)^2, \quad \text{and} \tag{6}$$

$$E(\bar{x}_i - a)^2 = (A - a^2)/s_i, \quad \text{therefore} \tag{7}$$

$$E\left(\sum_{i=1}^{N} s_i(\bar{x}_i - \bar{x})^2\right) = (N - 1)(A - a^2). \tag{8}$$

Similarly by using (1)

$$E\left(\sum_{i=1, j=1}^{N, s_i} (x_{ij} - \bar{x})^2\right) = (s - 1)(A - a^2). \tag{9}$$

Thus, in a statistically homogeneous set of items, classified as in (2), the following estimates of Variance have the same expected value:

$$\begin{aligned} V &= \frac{S}{s-1}, \quad &\text{where} \quad S &= \sum_{i=1, j=1}^{N, s_i} (x_{ij} - \bar{x})^2 \\ V_i &= \frac{S_i}{s-N}, \quad &\text{where} \quad S_i &= \sum_{i=1, j=1}^{N, s_i} (x_{ij} - \bar{x}_i)^2 \\ V_{\bar{x}} &= \frac{S_{\bar{x}}}{N-1}, \quad &\text{where} \quad S_{\bar{x}} &= \sum_{i=1}^{N} s_i(\bar{x}_i - \bar{x})^2. \end{aligned} \tag{10}$$

These estimates are used in applying the analysis of variance to the study of the correlation ratio, η, for uncorrelated material, where $\eta^2 = S_{\bar{x}}/S$.

OFFICE OF THE CHIEF SIGNAL OFFICER,
WASHINGTON, D. C.

[5] Rietz, H. L., loc. cit. p. 733.