ON THE ANALYSIS OF VARIANCE IN CASE OF MULTIPLE CLASSIFICATIONS WITH UNEQUAL CLASS FREQUENCIES

By Abraham Wald[1]

Columbia University

In a previous paper[2] the author considered the case of a single criterion of classification with unequal class frequencies and derived confidence limits for σ'^2/σ^2 where σ'^2 denotes the variance associated with the classification, and σ^2 denotes the residual variance. The scope of the present paper is to extend those results to the case of multiple classifications with unequal class frequencies.

For the sake of simplicity of notations we will derive the required confidence limits in the case of a two-way classification, the extension to multiple classifications being obvious.

Consider a two-way classification with p rows and q columns. Let y be the observed variable, and let n_{ij} be the number of observations in the ith row and jth column. Denote by $y_{ij}^{(k)}$ the kth observation on y in the ith row and jth column ($k = 1, \cdots, n_{ij}$). Let the total number of observations be N. We order the N observations and let y_α be the αth observation on y in that order. Consider the variables:

$$t, t_1, \cdots, t_p, v_1, \cdots, v_q,$$

and denote by t_α the αth observation on t, by $t_{i\alpha}$ the αth observation on t_i and by $v_{j\alpha}$ the αth observation on v_j. The values of t_α, $t_{i\alpha}$ and $v_{j\alpha}$ are defined as follows:

$$t_\alpha = 1 \ (\alpha = 1, \cdots, N),$$

$$t_{i\alpha} = 1 \text{ if } y_\alpha \text{ lies in the } i\text{th row},$$

$$t_{i\alpha} = 0 \text{ if } y_\alpha \text{ does not lie in the } i\text{th row},$$

$$v_{j\alpha} = 1 \text{ if } y_\alpha \text{ lies in the } j\text{th column},$$

$$v_{j\alpha} = 0 \text{ if } y_\alpha \text{ does not lie in the } j\text{th column}.$$

We make the assumptions

$$y_{ij}^{(k)} = x_{ij}^{(k)} + \epsilon_i + \eta_j,$$

where the variates $x_{ij}^{(k)}$, ϵ_i, η_j ($i = 1, \cdots, p; j = 1, \cdots, q; k = 1, \cdots, n_{ij}$) are independently and normally distributed, the variance of $x_{ij}^{(k)}$ is σ^2, the variance of ϵ_i is σ'^2, the variance of η_j is σ''^2, and the mean values of ϵ_i and η_j are zero.

---

[1] Research under a grant-in-aid from the Carnegie Corporation of New York.

[2] "A note on the analysis of variance with unequal class frequencies," *Annals of Math. Stat.*, Vol. 11 (1940).

Let the sample regression of y on $t, t_1, \cdots, t_{p-1}, v_1, \cdots, v_{q-1}$ be

$$Y = at + b_1t_1 + \cdots + b_{p-1}t_{p-1} + d_1v_1 + \cdots + d_{q-1}v_{q-1}.$$

We want to derive confidence limits for

$$\sigma'^2/\sigma^2 = \lambda^2.$$

Let us introduce the notations:

$$\begin{aligned}
\sum_\alpha t_\alpha t_{i\alpha} &= a_{0i} && (i = 1, \cdots, p-1),\\
\sum t_\alpha v_{j\alpha} &= a_{0p-1+j} && (j = 1, \cdots, q-1),\\
\sum t_{i\alpha} t_{j\alpha} &= a_{ij} && (i, j = 1, \cdots, p-1),\\
\sum t_{i\alpha} v_{j\alpha} &= a_{ip-1+j} && (i = 1, \cdots, p-1; j = 1, \cdots, q-1),\\
\sum v_{i\alpha} v_{j\alpha} &= a_{p-1+i\ p-1+j} && (i, j = 1, \cdots, q-1),\\
\| c_{ij} \| &= \| a_{ij} \|^{-1} && (i, j = 0, 1, \cdots, p+q-2).
\end{aligned}$$

Let the regression of $x_i^{(k)}$ on $t, t_1, \cdots, t_{p-1}, v_1, \cdots, v_{q-1}$ be

$$X = a^*t + b_1^*t_1 + \cdots + b_{p-1}^*t_{p-1} + d_1^*v_1 + \cdots + d_{q-1}^*v_{q-1}.$$

The regression of $\epsilon_i + \eta_j$ on the same independent variables is evidently equal to

$$\begin{aligned}
\epsilon_1t_1 + \cdots + \epsilon_pt_p + \eta_1v_1 + \cdots + \eta_qv_q \\
&= (\eta_q + \epsilon_p)t + (\epsilon_1 - \epsilon_p)t_1 + \cdots + (\epsilon_{p-1} - \epsilon_p)t_{p-1}\\
&\quad + (\eta_1 - \eta_q)v_1 + \cdots + (\eta_{q-1} - \eta_q)v_{q-1},
\end{aligned}$$

since $t_p = t - t_1 - \cdots - t_{p-1}$ and $v_q = t - v_1 - \cdots - v_{q-1}$. Hence

$$b_i = b_i^* + (\epsilon_i - \epsilon_p), \qquad (i = 1, \cdots, p-1), \tag{1}$$

and therefore

$$\begin{aligned}
\sigma_{b_ib_j} = \sigma_{b_i^*b_j^*} + \sigma_{(\epsilon_i-\epsilon_p)(\epsilon_j-\epsilon_p)} &= c_{ij}\sigma^2 + \sigma_{\epsilon_i\epsilon_j} + \sigma_{\epsilon_p\epsilon_p}\\
&= [c_{ij} + (1 + \delta_{ij})\lambda^2]\sigma^2, \quad (i, j = 1, \cdots, p-1),
\end{aligned} \tag{2}$$

where δ_{ij} is the Kronecker delta, i.e. $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ii} = 1$. Denote $c_{ij} + (1 + \delta_{ij})\lambda^2$ by c'_{ij}. Since the expected value of b_i^* is equal to zero, on account of (1) also the expected value of b_i is equal to zero. Let

$$\| g_{ij} \| = \| c'_{ij} \|^{-1}, \qquad (i, j = 1, \cdots, p-1).$$

Then

$$\frac{1}{\sigma^2} \sum_{j=1}^{p-1} \sum_{i=1}^{p-1} g_{ij} b_i b_j \tag{3}$$

has the χ^2-distribution with $p - 1$ degrees of freedom. The expression

$$\frac{1}{\sigma^2} \sum_{\alpha=1}^{N} (y_\alpha - Y_\alpha)^2, \tag{4}$$

has the χ^2-distribution $N - p - q + 1$ degrees of freedom. The expressions (3) and (4) are independently distributed. Hence

$$\frac{N - p - q + 1}{p - 1} \frac{\Sigma\Sigma g_{ij} b_i b_j}{\Sigma (y_\alpha - Y_\alpha)^2}, \tag{5}$$

has the F-distribution (analysis of variance distribution). We will now show that (5) is a monotonic function of λ^2. It is known that $\Sigma\Sigma g_{ij} b_i b_j$ is invariant under linear transformations, i.e.

$$\Sigma\Sigma g_{ij} b_i b_j = \Sigma\Sigma g'_{ij} b'_i b'_j ,$$

where b'_i is an arbitrary linear function, say $\mu_{i1} b_1 + \cdots + \mu_{ip-1} b_{p-1}$ of $b_1, \cdots, b_{p-1}$ $(i = 1, \cdots, p - 1)$ and

$$\| g'_{ij} \| = \| \sigma_{b'_i b'_j} \|^{-1}.$$

We can choose the matrix $\| \mu_{ij} \|$ such that

$$\epsilon'_i = \mu_{i1}(\epsilon_1 - \epsilon_p) + \cdots + \mu_{ip-1}(\epsilon_{p-1} - \epsilon_p), \qquad (i = 1, \cdots, p - 1),$$

are independently distributed and $\sigma^2_{\epsilon'_i} = \sigma'^2$. The coefficients μ_{ij} of course do not depend on σ'. We have

$$\sigma_{b'_i b'_j} = \sigma_{b^*_i b^*_j} + \delta_{ij} \sigma'^2, \qquad (\delta_{ij} = \text{Kronecker delta}).$$

Now let

$$b''_i = \nu_{i1} b'_1 + \cdots + \nu_{ip-1} b'_{p-1}, \qquad (\nu = 1, \cdots, p - 1),$$

where $\| \nu_{ij} \|$ is an orthogonal matrix and is chosen such that $b^{*''}_1, \cdots, b^{*''}_{p-1}$ are independently distributed. On account of the orthogonality of $\| \nu_{ij} \|$ we obviously have

$$\sigma^2_{b''_i} = \sigma^2_{b^{*''}_i} + \sigma'^2 ; \qquad \sigma_{b''_i b''_j} = 0 \qquad \text{for } i \neq j.$$

Hence

$$\sum \sum g_{ij} b_i b_j = \sum_{i=1}^{p-1} \frac{b''^2_i}{\sigma^2_{b^{*''}_i} + \lambda^2 \sigma^2}. \tag{6}$$

The right hand side of (6) is evidently a monotonic function of λ^2 which proves our statement. The endpoints of the confidence interval for λ^2 are the roots in λ^2 of the equations

$$\frac{N - p - q + 1}{p - 1} \frac{\Sigma\Sigma g_{ij} b_i b_j}{\Sigma (y_\alpha - Y_\alpha)^2} = F_2 ; \qquad \frac{N - p - q + 1}{p - 1} \frac{\Sigma\Sigma g_{ij} b_i b_j}{\Sigma (y_\alpha - Y_\alpha)^2} = F_1, \tag{7}$$

where F_2 denotes the upper, and F_1 the lower critical value of F.

The derivation of the required confidence limits in case of classifications in more than two ways can be carried out in the same way and I shall merely state here the results.

Consider r criterions of classifications and denote by p_u the number of classes in the uth classification ($u = 1, \cdots, r$). Denote by $n_{i_1 \cdots i_r}$ the number of observations which belong to the i_1th class of the first classification, i_2th class of the second classification, $\cdots$, and to the i_rth class of the rth classification. Let $y^{(k)}_{i_1 \cdots i_r}$ be the kth observation on y in the set of observations belonging to the classes mentioned above ($k = 1, \cdots, n_{i_1 \cdots i_r}$). We make the assumption

$$y^{(k)}_{i_1 \cdots i_r} = x^{(k)}_{i_1 \cdots i_r} + \epsilon^{(1)}_{i_1} + \cdots + \epsilon^{(r)}_{i_r},$$

where the variates

$$x^{(k)}_{i_1 \cdots i_r}, \ \epsilon^{(1)}_{i_1}, \cdots, \epsilon^{(r)}_{i_r} \quad (i_u = 1, \cdots, p_u; \ u = 1, \cdots, r; \ k = 1, \cdots, n_{i_1 \cdots i_r}),$$

are independently and normally distributed, the variance of $x^{(k)}_{i_1 \cdots i_r}$ is σ^2, the variance of $\epsilon^{(u)}_{i_u}$ is σ_u^2 and the mean value of $\epsilon^{(u)}_{i_u}$ is zero ($i_u = 1, \cdots, p_u$; $u = 1, \cdots, r$).

Let N be the total number of observations. We order the observations in a certain order and denote by y_α the αth observation in that order ($\alpha = 1, \cdots, N$). Consider the variables:

$$t, \ t^{(u)}_{i_u}, \qquad (u = 1, \cdots, r; \ i_u = 1, \cdots, p_u),$$

and denote by t_α the αth observation on t and by $t^{(u)}_{i_u \alpha}$ the αth observation on $t^{(u)}_{i_u}$. The values of t_α and $t^{(u)}_{i_u \alpha}$ are given as follows:

$t_\alpha = 1$ ($\alpha = 1, \cdots, N$),

$t^{(u)}_{i_u \alpha} = 1$ if y_α lies in the i_uth class of the uth classification,

$t^{(u)}_{i_u \alpha} = 0$ if y_α does not lie in the i_uth class of the uth classification.

Let the sample regression of y on $t, t^{(u)}_{i_u}$ be given by

$$Y = at + \sum_{u=1}^{r} \sum_{i_u=1}^{p_u - 1} b^{(u)}_{i_u} t^{(u)}_{i_u}.$$

Let the covariance of $b^{(u)}_{i_u}$ and $b^{(u)}_{j_u}$ be given by $C^{(u)}_{i_u j_u} \sigma^2$ under the assumption that $\sigma_1 = \sigma_2 = \cdots = \sigma_r = 0$. The matrix $\| C^{(u)}_{i_u j_u} \|$ ($i_u, j_u = 1, \cdots, p_u - 1$) can be calculated by known methods of the theory of least squares. Let

$$\| g^{(u)}_{i_u j_u} \| = \| C^{(u)}_{i_u j_u} + (1 + \delta_{i_u j_u}) \lambda_u^2 \|^{-1} \qquad (i_u, j_u = 1, \cdots, p_u - 1),$$

where $\delta_{i_u j_u}$ is the Kronecker delta and $\lambda_u^2 = \sigma_u^2 / \sigma^2$. Then the lower and upper confidence limits for λ_u^2 are given by the roots in λ_u^2 of the equations

$$\frac{N - \sum_{u=1}^{r} p_u + r - 1}{p_u - 1} \cdot \frac{\sum_{i_u=1}^{p_u - 1} \sum_{j_u=1}^{p_u - 1} g^{(u)}_{i_u j_u} b^{(u)}_{i_u} b^{(u)}_{j_u}}{\sum_{\alpha=1}^{N} (y_\alpha - Y_\alpha)^2} = F_i \qquad (i = 1, 2), \tag{8}$$

where F_2 is the upper and F_1 the lower critical value of the analysis of variance distribution with $p_u - 1$ and $N - \sum_{u=1}^{r} p_u + r - 1$ degrees of freedom. In case of a single criterion of classification the confidence limits (8) are identical with those given in my previous paper.

THE FREQUENCY DISTRIBUTION OF A GENERAL MATCHING PROBLEM

By T. N. E. Greville

Bureau of the Census

1. Introduction. This paper considers the matching of two decks of cards of arbitrary composition, and the complete frequency distribution of correct matchings is obtained, thus solving a problem proposed by Stevens.[1] It is also shown that the results can be interpreted in terms of a contingency table.

Generalizing a problem considered by Greenwood,[2] let us consider the matching of two decks of cards consisting of t distinct kinds, all the cards of each kind being identical. The first or "call" deck will be composed of i_1 cards of the first kind, i_2 of the second, etc., such that

$$i_1 + i_2 + i_3 + \cdots + i_t = n;$$

and the second or "target" deck will contain j_1 cards of the first kind, j_2 of the second, etc., such that

$$j_1 + j_2 + \cdots + j_t = n.$$

Any of the i's or j's may be zero. It is desired to calculate, for a given arrangement of the "call" deck, the number of possible arrangements of the "target" deck which will produce exactly r matchings between them ($r = 0, 1, 2, \cdots, n$). It is clear that these frequencies are independent of the arrangement of the call deck. For convenience the call deck may be thought of as arranged so that all the cards of the first kind come first, followed by all those of the second kind, and so on.

2. Formulae for the frequencies. Let us consider the number of arrangements of the target deck which will match the cards in the k_1th, k_2th, $\cdots$, k_sth positions in the call deck, regardless of whether or not matchings occur elsewhere. Let the cards in these s positions in the call deck consist of c_1 of the first kind, c_2 of the second, etc. Then:

$$c_1 + c_2 + \cdots + c_t = s.$$

The number of such arrangements of the target deck is

$$\frac{(n-s)!}{\prod_{h=1}^{t} (j_h - c_h)!}. \tag{1}$$

[1] W. L. Stevens, *Annals of Eugenics*, Vol. 8 (1937), pp. 238–244.

[2] J. A. Greenwood, *Annals of Math. Stat.*, Vol. 9 (1938), pp. 56–59.