# ON SOME CLASSIFICATION STATISTICS

*By* S. JOHN

*Indian Statistical Institute, Calcutta*

*SUMMARY.* Various authors have suggested various statistics for use in classification problems. To set up the classification procedure and to study the performance characteristics of procedures thus set up, it is necessary to know the distributions of the statistics used. This paper gives the exact distributions of several classification statistics.

## 1. Introduction

Research workers are often faced with the problem of assigning an individual, known to belong to one or other of two classes, to its correct class. They do this by observing various characteristics of the individual. Some characteristics whose ranges for the two classes are completely disjoint would have been ideal for classification purposes. But it is generally the case that there is a certain range of values where individuals from both the classes may appear. This prevents us from being always correct. But if we know the distributions of the characteristics in both the classes, we can devise a procedure in which the chances of misclassification are least. This will be achieved if we use the likelihood ratio as a criterion. The well-known procedure of using the discriminant function is equivalent to the one using the likelihood ratio provided the distributions of the characteristics are in both the populations, multivariate normal with identical dispersion matrices. If we denote the characteristics observed by $x = (x_1, ..., x_p)$ the common dispersion matrix by $\Sigma$ and the mean vectors in the two populations $P_1$ and $P_2$ by $\mu^{(1)}$ and $\mu^{(2)}$, the discriminant function is

$$D(x) = (\mu^{(2)} - \mu^{(1)})\Sigma^{-1}x'. \qquad \dots \quad (1.1)$$

Individuals with small $D(x)$ values are assigned to $P_1$ and those with large $D(x)$ values are assigned to $P_2$.

If the distributions of $x$ in $P_1$ and $P_2$ are unknown, we are unable to use this procedure. To meet this situation, Wald (1944) suggested that we could use the function $U(x)$ obtained by substituting in $D(x)$ sample estimates of $\mu^{(1)}$, $\mu^{(2)}$ and $\Sigma$. To set up the classification procedure and to evaluate the chances of misclassification involved in adopting the procedure thus set up, we require the distributions of $U(x)$ for $x$'s from $P_1$ and $P_2$. Wald tackled this problem in 1944. He exhibited $U(x)$ as a function of three statistics $m_1$, $m_2$, $m_3$ whose joint distribution he expressed as a product of three factors, the first an exponential factor, the second a product of gamma and beta functions and the third an expectation of a power of a random determinant. In 1951, Anderson evaluated this expectation for the special case where $\mu^{(1)}$ and $\mu^{(2)}$ are parallel vectors. Later Sitgreaves (1952) gave an alternate derivation of the joint distribution of $m_1$, $m_2$ and $m_3$, again for the same special case as considered by Anderson. Still it was only the joint distribution of $m_1$, $m_2$ and $m_3$ and not the distribution of Wald's classification statistic itself. Harter (1951) derived the exact distribution for the univariate case under the assumption that at least one of the populations has zero mean. For the multivariate case he showed that an approximation could be derived provided the mean vector of one of the populations is zero and the dispersion matrix is estimated with a large even degree of freedom.

In a previous paper (John, 1959), the author obtained the exact distribution of Wald's classification statistic for the case where the dispersion matrix is known. Unfortunately, in that paper the assumption was made that there was no loss of generality in assuming $\mu^{(1)} = 0$. In this paper, we remove this restriction.

Since Wald made his suggestion other authors have come forward with other procedures. Of these, the one suggested by Rao (1954) was specially constructed so as to be most powerful in discriminating between alternatives which were close to each other. But in no case was the necessary distribution available. We obtain these distributions. Some remarks on the relative merits of these statistics are also included.

## 2. NOTATION

We shall denote the two alternate populations by $P_1$ and $P_2$. It is assumed that the classification procedure is based on $p$ characteristics $x_1, x_2, \ldots, x_p$. In what follows, $x = (x_1, \ldots, x_p)$ will be supposed to follow the multivariate normal law in both $P_1$ and $P_2$. We shall assume that these two distributions have the same dispersion matrix $\Sigma$ but different mean vectors, $\mu^{(1)} = (\mu_1^{(1)}, \ldots, \mu_p^{(1)})$ for $P_1$ and $\mu^{(2)} = (\mu_1^{(2)}, \ldots, \mu_p^{(2)})$ for $P_2$.

Estimates of $\mu^{(1)}$ and $\mu^{(2)}$ are required later. (We are, in this paper, assuming that $\Sigma$ is known). If $x_{i\alpha}^{(1)}$ $(\alpha = 1, 2, \ldots, N_1;\ i = 1, 2, \ldots, p)$ is random sample of $N_1$ individuals from $P_1$ and $x_{i\alpha}^{(2)}$ $(\alpha = 1, \ldots, N_2;\ i = 1, 2, \ldots, p)$ is a sample of $N_2$ individuals from $P_2$,

$$\bar{x}^{(1)} = (\bar{x}_1^{(1)}, \bar{x}_2^{(1)}, \ldots, \bar{x}_p^{(1)})$$

and

$$\bar{x}^{(2)} = (\bar{x}_1^{(2)}, \bar{x}_2^{(2)}, \ldots, \bar{x}_p^{(2)})$$

where

$$\bar{x}_i^{(1)} = \frac{\sum\limits_{\alpha=1}^{N_1} x_{i\alpha}^{(1)}}{N_1} \text{ and } \bar{x}_i^{(2)} = \frac{\sum\limits_{\alpha=1}^{N_2} x_{i\alpha}^{(2)}}{N_2}$$

can be used as estimates of $\mu^{(1)}$ and $\mu^{(2)}$ respectively.

## 3. Wald's classification statistic

In the notation developed above, the statistic suggested by Wald is

$$U(x) = (\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}x'. \qquad \dots \quad (3.1)$$

Using this statistic we may set up a classification procedure as follows: Determine a constant $d_1$ such that

$$Pr\,(U(x) < d_1 | x \,\epsilon\, P_1) = \alpha_1 \qquad \dots \quad (3.2)$$

and a constant $d_2$ such that

$$Pr\,(U(x) < d_2 | x \,\epsilon\, P_2) = \alpha_2. \qquad \dots \quad (3.3)$$

What $\alpha_1$ and $x_2$ are—will be made clear below. Assign the individual to $P_1$ or $P_2$ according as

$$U(x) \begin{array}{l} < \min(d_1, d_2) \\ \geqslant \max(d_1, d_2) \end{array} \qquad \dots \quad (3.4)$$

If this classification procedure is adopted, the chance of misclassifying an individual from $P_1$ will not exceed $\alpha_1$; so also the chance of misclassifying an individual from $P_2$ will not exceed $\alpha_2$.

If $d_1 < d_2$ and $d_1 \leqslant U(x) \leqslant d_2$, or if $d_2 < d_1$ and $d_2 \leqslant U(x) \leqslant d_1$, we will not classify the individual as belonging to $P_1$ or $P_2$. In the first case the individuals who are unclassified are those about whom there is a suspicion that they came neither from $P_1$ nor from $P_2$ but from some other population. In the second case the individuals who are unclassified are those who are claimed by both $P_1$ and $P_2$.

If it is preferred not to have a region of indeterminacy, we may adopt the procedure of classifying the individual as belonging to $P_1$ or $P_2$ according as

$$U(x) \lessgtr U(\tfrac{1}{2}[\bar{x}^{(1)}+\bar{x}^{(2)}]). \qquad \dots \quad (3.5)$$

The chances of errors of the two kinds involved in this procedure can be assessed with the help of the distributions which we obtain below.

*The distribution of Wald's classification statistic.* The distribution of the statistic $(\bar{x}^{(2)}-\bar{x}^{(1)})\ \Sigma^{-1}x'$ is required. We shall obtain the distribution of this statistic under the assumption that $x$ came from $P_1$: its distribution for $x$'s from $P_2$ can be obtained similarly.

Rather than $(\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}x'$ we prefer to consider the statistic

$$T = 2\,\frac{(N_1N_2)^{\frac{1}{2}}}{(N_1+N_2)^{\frac{1}{2}}}\,(\bar{x}^{(2)}-\bar{x}^{(1)})\ \Sigma^{-1}x'. \qquad \dots \quad (3.6)$$

Put $$u = xA,\ v = \frac{(N_1N_2)^{\frac{1}{2}}}{(N_1+N_2)^{\frac{1}{2}}}\ (\bar{x}^{(2)}-\bar{x}^{(1)})A \qquad \ldots \quad (3.7)$$

where $A$ is a $(p\times p)$ matrix such that $A'\Sigma A = I$.

Then

$$T = 2vu' = 2\sum_{i=1}^{p} u_i v_i = \tfrac{1}{2}\sum_{i=1}^{p}(v_i+u_i)^2-\tfrac{1}{2}\sum_{i=1}^{p}(v_i-u_i)^2 = t-w \text{ (say)}. \qquad \ldots \quad (3.8)$$

$t$ and $w$ have independent noncentral chi-square distributions, in each case with degree of freedom $p$ but with noncentralities given respectively by

$$\lambda_1 = \tfrac{1}{2}\, c\, \Sigma^{-1}c' \quad \text{and} \quad \lambda_2 = \tfrac{1}{2}\, d\, \Sigma^{-1}d' \qquad \ldots \quad (3.9)$$

where $$c = \mu^{(1)}+(N_1^{-1}+N_2^{-1})^{-\frac{1}{2}}(\mu^{(2)}-\mu^{(1)}) \qquad \ldots \quad (3.10)$$

and $$d = \mu^{(1)}-(N_1^{-1}+N_2^{-1})^{-\frac{1}{2}}(\mu^{(2)}-\mu^{(1)}) \qquad \ldots \quad (3.11)$$

The joint density of $t$ and $w$ is

$$f(t, w) = 2^{-p}\, e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\,\lambda_2^s}{r\,!\,s\,!\,\Gamma\left(\frac{p+2r}{2.}\right)\Gamma\left(\frac{p+2s}{2}\right)}\, t^{(p+2r-2)/2}\, w^{(p+2s-2)/2}\, e^{-(t+w)/2}. \qquad \ldots \quad (3.12)$$

Hence the density function of $T$ is given by

$$g(T)dT = \int\ldots\int_{T<t-w<T+dT} f(t, w)dtdw. \qquad \ldots \quad (3.13)$$

On simplification this gives

$$g(T) = 2^{-p}\, e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\,\lambda_2^s}{r\,!\,s\,!\,\Gamma\left(\frac{p+2r}{2}\right)}\, T^{(r+s+p-2)/2}\, W_{l,m}(T)(T>0) \qquad \ldots \quad (3.14)$$

where $$\cdot l = \frac{r-s}{2}\ \ m = \frac{r+s+p-1}{2} \qquad \ldots \quad (3.15)$$

and where $W_{l,m}(T)$ is Whittaker's confluent hypergeometric function defined by

$$W_{l,m}(x) = \frac{x^{m+\frac{1}{2}}}{\Gamma(m-l+\frac{1}{2})}e^{-x/2}\int_0^{\infty} e^{-tx}\, t^{m-l-\frac{1}{2}}\,(1+t)^{m+l-\frac{1}{2}}\, dt. \qquad \ldots \quad (3.16)$$

If $T<0$,

$$g(T) = 2^{-p}\, e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\lambda_2^s}{r!\ \ s!\ \Gamma\left(\frac{p+2s}{2}\right)}\ (-T)^{(r+s+p-2)/2}\, W_{-l,m}(-T) \qquad \ldots \quad (3.17)$$

## 4. Anderson's classification statistic

In this section we consider a statistic suggested by Anderson for classification purposes. In our notation this statistic is

$$Q \equiv (\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}x'-\tfrac{1}{2}(\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}(\bar{x}^{(2)}+\bar{x}^{(1)})'. \quad \dots \quad (4.1)$$

The steps involved in setting up the classification procedure are much the same as for Wald's classification statistic. We determine two constants $c_1$ and $c_2$ such that

$$Pr(Q > c_1 | x \epsilon P_1) = \alpha_1 \quad \dots \quad (4.2)$$

and

$$Pr(Q < c_2 | x \epsilon P_2) = \alpha_2 \quad \dots \quad (4.3)$$

$\alpha_1$ and $\alpha_2$ being respectively the levels at which the two kinds of error are to be controlled. Individuals are assigned to $P_1$ or $P_2$ according as

$$Q \begin{array}{l} < \min(c_1, c_2) \\ \geqslant \max(c_1, c_2). \end{array} \quad \dots \quad (4.4)$$

If this procedure is adopted the probability of wrongly assigning to $P_2$ an individual from $P_1$ will not exceed $\alpha_1$; so also the probability of wrongly assigning to $P_1$ an individual from $P_2$ will not exceed $\alpha_2$.

It is to be noted that we are abstaining from classifying individuals for whom

$$\min(c_1, c_2) \leqslant Q(x) \leqslant \max(c_1, c_2). \quad \dots \quad (4.5)$$

These are again individuals who are either claimed by both $P_1$ and $P_2$ or individuals about whom there is a suspicion that they came neither from $P_1$ nor from $P_2$ but from some other population.

If it is desired not to have a region of indecision the following classification procedure may be followed. Assign individuals to $P_1$ or $P_2$ according as

$$Q \lesseqgtr 0. \quad \dots \quad (4.6)$$

It is not difficult to see that this procedure is equivalent to assigning individuals to $P_1$ or $P_2$ according as

$$U(x) \lesseqgtr U(\tfrac{1}{2}[\bar{x}^{(1)}+\bar{x}^{(2)}]). \quad \dots \quad (4.7)$$

In whatever way the classification procedure is set up, if it is based on $Q$, the chances of errors of the two kinds involved in such a procedure can be evaluated using the distribution of $Q$ obtained below. The author has given elsewhere simple approximations as well as exact expressions for these probabilities.

*Distribution of Anderson's classification statistic.* We shall now obtain the distribution of Anderson's classification statistic for $x$'s from $P_1$; its distribution for $x$'s from $P_2$ be obtained similarly.

Since the steps involved in arriving at these distributions are similar to those for Wald's classification statistic, only the final expressions will be given.

$$z = 4N_1N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}Q \qquad \dots \quad (4.8)$$

has the density function

$$f_1(z) = \begin{cases} e^{-z(a-b)/4} \sum\limits_{r=0}^{\infty} \sum\limits_{s=0}^{\infty} c_{rs} \dfrac{a^{\frac{1}{2}(p+2r)} b^{\frac{1}{2}(p+2s)}}{\Gamma\left(\dfrac{p+2r}{2}\right)} z^{\frac{1}{2}(p+r+s)-1} W_{l,m}(cz) \; (z > 0) \\[2ex] e^{-z(a-b)/4} \sum\limits_{r=0}^{\infty} \sum\limits_{s=0}^{\infty} c_{rs} \dfrac{a^{\frac{1}{2}(p+2r)} b^{\frac{1}{2}(p+2s)}}{\Gamma\left(\dfrac{p+2s}{2}\right)} (-z)^{\frac{1}{2}(p+r+s)-1} W_{-l,m}(-cz) \; (z < 0) \end{cases} \qquad \dots \quad (4.9)$$

where

$$a = (1+\rho)^{-1}, \quad b = (1-\rho)^{-1}, \quad c = (a+b)/2 \qquad \dots \quad (4.10)$$

$$\rho = (N_2-N_1)(N_1+N_2)^{-\frac{1}{2}}[4N_1N_2+N_1+N_2]^{-\frac{1}{2}} \qquad \dots \quad (4.11)$$

$$c_{rs} = e^{-\lambda_1-\lambda_2} \frac{\lambda_1^r\lambda_2^s}{r!s!} (4c)^{-(p+r+s)/2} \qquad \dots \quad (4.12)$$

$$\lambda_1 = \tfrac{1}{2}(1+\rho)^{-1}N_1N_2[(N_1+N_2)^{-\frac{1}{2}}-(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}]^2\delta^2 \qquad \dots \quad (4.13)$$

$$\lambda_2 = \tfrac{1}{2}(1-\rho)^{-1}N_1N_2[(N_1+N_2)^{-\frac{1}{2}}+(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}]^2\delta^2 \qquad \dots \quad (4.14)$$

$$\delta^2 = (\mu^{(2)}-\mu^{(1)})\Sigma^{-1}(\mu^{(2)}-\mu^{(1)})'. \qquad \dots \quad (4.15)$$

$l$ and $m$ have the same meanings as before.

It is easy to see that if $N_1 = N_2$, $z$ has the same distribution as $T$ with $\lambda_1$ and $\lambda_2$ given by equations (4.13) and (4.14).

## 5. Rao's statistic

Rao (1954) proposes the following statistic for classification purposes

$$R = \frac{a_1^2-a_2^2}{b_2^2} U\Sigma^{-1}U' + \frac{2(a_1-a_2)}{b_1b_2} U\Sigma^{-1}T' \qquad \dots \quad (5.1)$$

where

$$a_1 = N_2(N_1+N_2)^{-1},$$
$$a_2 = -N_1(N_1+N_2)^{-1}$$
$$b_1 = N_1^{-1}+N_2^{-1},$$
$$b_2 = 1+(N_1+N_2)^{-1}$$
$$U = x-(N_1\bar{x}^{(1)}+N_2\bar{x}^{(2)})(N_1+N_2)^{-1}$$
$$T = \bar{x}^{(1)}-\bar{x}^{(2)}.$$

When $N_1 = N_2$, this is equivalent to Anderson's statistic.

For later use we define

$$d = 2b_1^{-1}\, b_2(a_1+a_2)^{-1}. \qquad \dots \quad (5.2)$$

Assuming that $x$ came from $P_1$,

$$R_s = \frac{b_2^2}{a_1^2-a_2^2}[1+(N_1+N_2)^{-1}]^{-\frac{1}{2}}[1+(N_1+N_2)^{-1}+b_1d^2]^{-\frac{1}{2}}R \qquad \dots \quad (5.3)$$

has the same distribution as $z$, with

$$\lambda_1 = \tfrac{1}{2}(1+\rho)^{-1}[N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+1)^{-\frac{1}{2}}+ \\ +\{1+(N_1+N_2)^{-1}+b_1d^2\}^{-\frac{1}{2}}(a_1+d)]^2\delta^2 \qquad \dots \quad (5.4)$$

$$\lambda_2 = \tfrac{1}{2}(1-\rho)^{-1}[N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+1)^{-\frac{1}{2}}- \\ -\{1+(N_1+N_2)^{-1}+b_1d^2\}^{-\frac{1}{2}}(a_1+d)]^2\delta^2 \qquad \dots \quad (5.5)$$

$$\rho = [1+(N_1+N_2)^{-1}]^{\frac{1}{2}}[1+(N_1+N_2)^{-1}+b_1d^2]^{-\frac{1}{2}}. \qquad \dots \quad (5.6)$$

The changes to be made to obtain the distribution of this statistic for $x$'s from $P_2$ are obvious.

The distributions we have obtained are found to contain certain parameters. The distributions of Anderson's classification statistic and Rao's classification statistic involve one parameter each. In either case it is the distance between the two populations $P_1$ and $P_2$. On the other hand the distribution of Wald's classification statistic contains two parameters. They are not simply related to the distance between $P_1$ and $P_2$. This feature makes Wald's classification statistic less attractive than either Anderson's statistic or Rao's statistic.

That the classification procedure using Wald's statistic is not invariant under changes of origin is another point on its debit side.

Since the distributions involve parameters, how shall we find constants $c_1$, $c_2$, $d_1$, $d_2$ etc? Consider Anderson's statistic and Rao's statistic. The parameter appearing in the distributions of either of these statistics is the distance $\delta$ between the two populations $P_1$ and $P_2$. We may require that if $\delta \geqslant \delta_0$ the probability of an error of the first kind, (i.e. of wrongly classifying an individual from $P_1$ as belonging to $P_2$) shall not exceed $\alpha_1$ and also that the probability of an error of the second kind (i.e. of wrongly classifying an individual from $P_2$ as belonging to $P_1$) shall not exceed $\alpha_2$. This requirement will be satisfied if we determine the constants $c_1$, $c_2$ etc., assuming that $\delta = \delta_0$. For a classification procedure thus set up, the probability of errors of the two kinds for other values of $\delta$ can be evaluated with the help of the distributions obtained in this paper.

A discussion of further problems of a similar nature will be included in a later paper.

I wish to thank Dr. C. R. Rao for his careful perusal of this paper.

### References

Anderson, T. W. (1951) : Classification by multivariate analysis. *Psychometrika*, **16**, 31.

Harter, H. L. (1951) : On the distribibution of Wald's classification statistic. *Ann. Math. Stat.*, **22**, 58.

John, S. (1959) : The distribution of Wald's classification statistic when the dispersion matrix is known. *Sankhyā*, **21**, 371.

Rao, C. R. (1954) : On a general theory of discrimination when the information on alternate hypotheses is based on samples. *Ann. Math. Stat.*, **25**, 651.

Sitgreaves, R. (1952) : On the distribution of two random matrices used in classification procedures. *Ann. Math. Stat.*, **23**, 263.

Wald, A. (1944) : On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.*, **15**, 145.

*Paper received : November, 1959.*