# SOME TECHNIQUES FOR SIMPLE CLASSIFICATION

CARL F. KOSSACK

UNIVERSITY OF OREGON

## 1. Introduction

In 1944 Wald[1] considered the problem of classifying a single multivariate observation, $z$, into one of two normally distributed parent populations, $\pi_1$ and $\pi_2$, when the only information available about the populations is contained in two samples of sizes $N_1$ and $N_2$, one drawn from each population. In order to obtain a classification technique, Wald assumed that the populations $\pi_1$ and $\pi_2$ have the same covariance matrix but unequal means and used the Neyman-Pearson[2] most powerful test for the hypothesis that $z$ belongs to $\pi_1$ against the single alternative hypothesis that $z$ belongs to $\pi_2$. The most powerful test for this hypothesis is given by the critical region $U \geqq d$, where $U = \sum_j \sum_i \sigma^{ij} z_i(\nu_j - \mu_j)$ and $\| \sigma^{ij} \|$ denotes the inverse matrix of the covariance matrix $\| \sigma_{ij} \|$, $z_i$ the $i$th variate of the single observation, $\nu_j$ and $\mu_j$ the means of the $j$th variate for the populations $\pi_1$ and $\pi_2$. The critical region $U \geqq d$ is then approximated by $R \geqq d$, where $R$ is the statistic obtained from $U$ by replacing $\sigma^{ij}$, $\nu_j$, and $\mu_j$ by their optimum estimates obtained from the two samples. In order to determine $d$ corresponding to a given probability of an error of the first kind (classifying $z$ in $\pi_2$ when $z$ belongs to $\pi_1$) and the associated probabilty of an error of the second kind (classifying $z$ in $\pi_1$ when $z$ belongs to $\pi_2$) for the case when $N_1$ and $N_2$ are large, Wald used the fact that $R$ can be approximated by means of the normal curve with means and covariance matrix obtained from the two samples.

In this paper we shall consider the problem of classifying an observation of a single variate into one of two normally distributed populations where the assumption of equal variances need not necessarily be valid. We shall distinguish this single-variate problem from the multivariate one by referring to it as simple classification.

## 2. Statement of the problem

We consider two variates $x$ and $y$ and assume that each is normally distributed and that each is independent of the other. A sample of size $N_1$ is drawn from the population $\pi_1$, the $x$-population, and a sample of size $N_2$ from the population $\pi_2$, the $y$-population. Denote by $x_i$ the $i$th observation on $x$ ($i = 1, 2, \cdots, N_1$) and by $y_j$ the $j$th observation on $y$ ($j = 1, 2, \cdots, N_2$). Denote by

[1] Abraham Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Annals of Math. Stat.*, vol. 15 (June, 1944).

[2] J. Neyman and E. S. Pearson, "Contributions to the theory of testing statistical hypotheses," *Stat. Res. Mem.*, vol. 1 (London, 1936).

$\nu_1$, $\sigma_1$ and $\nu_2$, $\sigma_2$ the mean and standard deviation respectively of $\pi_1$ and $\pi_2$. Let $z$ be a single observation where it is known a priori that $z$ has been drawn from either $\pi_1$ or $\pi_2$.

Our problem is to test the hypothesis $H_1$ that the population from which $z$ was drawn was $\pi_1$ on the basis of the observations $x_i$, $y_j$, and $z$ $(i = 1, 2, \cdots, N_1)(j = 1, 2, \cdots, N_2)$.

## 3. The statistic to be used for testing the hypothesis $\mathbf{H_1}$

Neyman and Pearson[3] have shown that in the case of testing a hypothesis $H_1$ against a single alternative hypothesis $H_2$, the critical region that is most powerful is given by the inequality

$$\frac{p_2(z)}{p_1(z)} \geqq k,$$

where $p_i(z)$ denotes the probability of $z$ under the hypothesis $H_i$, and $k$ is a constant determined so that the critical region should have the required size. The critical region to be used in our test of the hypothesis $H_1$ ($z$ belongs to $\pi_1$) against the alternative hypothesis $H_2$ ($z$ belongs to $\pi_2$) depends upon the assumption that we make about the parameters $\nu_1$, $\sigma_1$, and $\nu_2$, $\sigma_2$. We shall consider three cases:

(i) $\sigma_1 = \sigma_2 = \sigma, \qquad \nu_1 \neq \nu_2;$

(ii) $\nu_1 = \nu_2 = \nu, \qquad \sigma_1 \neq \sigma_2;$

(iii) $\nu_1 \neq \nu_2, \qquad \sigma_1 \neq \sigma_2.$

In each case we shall follow the Wald procedure of approximating the critical region arising out of these assumptions by using in place of the $\nu$'s and $\sigma$'s their optimum estimates obtained from the two samples.

Case (i). $\sigma_1 = \sigma_2 = \sigma, \qquad \nu_1 \neq \nu_2.$

In this case

$$p_1(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(z-\nu_1)^2}{2\sigma^2}}$$

$$p_2(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(z-\nu_2)^2}{2\sigma^2}}$$

The critical region is given by the inequality

$$\frac{p_2}{p_1} = e^{\frac{(z-\nu_1)^2 - (z-\nu_2)^2}{2\sigma^2}} \geqq k,$$

[3] *Op. cit.*

or, if $\nu_1 < \nu_2$, the inequality reduces to

$$z \geqq d. \tag{1}$$

In order to determine $d$ and the probability of making errors of type I and type II, we use the optimum estimates of the parameters $\nu_1$, $\nu_2$, and $\sigma^2$, as obtained from the two samples, namely,

$$\bar{x} = \frac{\sum_{i=1}^{N_1} x_i}{N_1}, \qquad \bar{y} = \frac{\sum_{j=1}^{N_2} y_j}{N_2},$$

$$s^2 = \frac{\sum_{i=1}^{N_1}(x_i - \bar{x})^2 + \sum_{j=1}^{N_2}(y_j - \bar{y})^2}{N_1 + N_2 - 2}$$

If we set $d = \bar{x} + \lambda s$, we would have

$$P(z \geqq \bar{x} + \lambda s \mid z\, c\, \pi_1) = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2} dt = P_{\text{I}}, \tag{2}$$

and

$$P(z < \bar{x} + \lambda s \mid z c \pi_2) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\frac{\bar{x} + \lambda s - \bar{y}}{s}} e^{-t^2/2}\, dt = P_{\text{II}}. \tag{3}$$

Thus, if we assign a value to $P_{\text{I}}$, then equation (2) is used to evaluate $\lambda$; and equation (3) determines the value of $P_{\text{II}}$ corresponding to that value of $\lambda$. The efficiency of this type of classification may be measured by the probability $P = 1 - P_{\text{I}} = 1 - P_{\text{II}}$. In order to determine the value of $\lambda$ corresponding to the case $P_{\text{I}} = P_{\text{II}}$, we have from the symmetry of the normal curve

$$\lambda = -\frac{\bar{x} + \lambda s - \bar{y},}{s}$$

or

$$\lambda = \frac{\bar{x} - \bar{y},}{2s}$$

and the corresponding critical region is given by

$$z \geqq \bar{x} + \lambda s = \frac{\bar{x} + \bar{y},}{2}$$

Wald's multivariate case would simplify to give as the critical region associated with this problem,

$$R = \frac{1}{s^2}(\bar{y} - \bar{x})\, z \geqq d.$$

By using the fact that the distribution of $R$ can be approximated by the normal distribution with mean value $\bar{\alpha}_1 = \frac{1}{s^2}\bar{x}(\bar{y} - \bar{x})$ if $H_1$ and mean value $\bar{\alpha}_2 = \frac{1}{s^2}\bar{y}(\bar{y} - \bar{x})$ if $H_2$ and standard deviation $\bar{s} = \frac{1}{s}(\bar{y} - \bar{x})$, the values of $d$, $P_{\text{I}}$, and $P_{\text{II}}$ corresponding to this critical region can be determined. The Wald critical region is the same as the critical region given by (3), since if

$$\frac{1}{s^2}(\bar{y} - \bar{x})z \geqq \frac{1}{s^2}\bar{x}(\bar{y} - \bar{x}) + \lambda\frac{1}{s}(\bar{y} - \bar{x}),$$

then $z \geqq \bar{x} + \lambda s$.

The efficiency of this classification technique can be judged by means of table 1 below.

TABLE 1

VALUES OF $P_{\text{I}}$ AND $P_{\text{II}}$ FOR THE CRITICAL REGION $z \geqq d$ CORRESPONDING TO VARYING DIFFERENCES IN MEANS

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | |
|---|---|---|---|---|---|
| | $\bar{x} - \bar{y} = s$ | $\bar{x} - \bar{y} = 2s$ | $\bar{x} - \bar{y} = 3s$ | $\bar{x} - \bar{y} = 4s$ | $\bar{x} - \bar{y} = 5s$ |
| .01 | .91 | .63 | .25 | .05 | .01 |
| .05 | .74 | .36 | .09 | .01 | ... |
| .10 | .61 | .24 | .04 | ... | ... |
| .20 | .44 | .12 | .02 | ... | ... |
| .50 | .16 | .02 | ... | ... | ... |
| .70 | .06 | .01 | ... | ... | ... |
| .90 | .01 | ... | ... | ... | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .69 | .84 | .93 | .98 | .99 |

Case (ii). $\nu_1 = \nu_2 = \nu, \qquad \sigma_1 \neq \sigma_2$.
The critical region in this case is

$$\frac{p_2}{p_1} = \frac{\sigma_1}{\sigma_2}e^{\frac{(z-\nu)^2}{2}\left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right]} \geqq k,$$

or, if $\sigma_2^2 > \sigma_1^2$,

$$(z - \nu)^2 \geqq d^2. \tag{4}$$

Let

$$m = \frac{\sum_{i=1}^{N_1} x_i + \sum_{j=1}^{N_2} y_j}{N_1 + N_2},$$

$$s_1^2 = \frac{\sum_{i=1}^{N_1}(x_i - m)^2}{N_1 - 1}, \qquad s_2^2 = \frac{\sum_{j=1}^{N_2}(y_j - m)^2}{N_2 - 1}.$$

We shall then take for our critical region

$$(z - m)^2 \geqq d^2,$$

and assume that the sampling distribution of $(z - m)^2$ can be approximated by the $\chi^2$ distribution. That is,

$$p(\chi^2) = c(\chi^2)^{\frac{1}{2}f-1}e^{-\frac{1}{2}\chi^2},$$

where

$$(z - m)^2 = s_1^2\,\chi^2, \qquad f = 1, \quad \text{if } H_1,$$

and if $s_2^2 = h^2 s_1^2$,

$$(z - m)^2 = h^2\,s_1^2,\,\chi^2, \qquad f = 1, \quad \text{if } H_2.$$

In order to evaluate $d^2$, $P_{\text{I}}$, and $P_{\text{II}}$, we have the following relationships:[4]

$$P\,[(z - m)^2 \geqq d^2 | H_1] = \int_{x_0^2}^{\infty} p(\chi^2)d\chi^2 = P_{\text{I}}, \tag{5}$$

and

$$P\,[(z - m)^2 < d^2 | H_2] = \int_{0}^{h^{-2}x_0^2} p(\chi^2)d\chi^2 = P_{\text{II}}. \tag{6}$$

For a given value of $P_{\text{I}}$, equation (5) determines the corresponding value of $x_0^2$ which yields the value of $d^2$ to use in the critical region from the relationship $d^2 = s_1^2\,x_0^2$. Relationship (5) would then be used to determine the corresponding value of $\text{P}_{\text{II}}$.

The efficiency of this method of classification can be judged from table 2 below.

TABLE 2

VALUES OF $P_{\text{I}}$ AND $P_{\text{II}}$ FOR THE CRITICAL REGION $(z - m)^2 \geqq d^2$ CORRESPONDING TO VARYING DIFFERENCES IN STANDARD DEVIATION

$[s_2 = hs_1, (h > 1)]$

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $h = 2$ | $h = 3$ | $h = 4$ | $h = 5$ | $h = 10$ | $h = 20$ | $h = 50$ | $h = 100$ |
| .01 | .80 | .59 | .48 | .37 | .20 | .10 | .05 | .02 |
| .05 | .66 | .48 | .36 | .30 | .15 | .07 | .04 | .01 |
| .10 | .57 | .40 | .31 | .25 | .12 | .06 | .03 | .01 |
| .20 | .48 | .32 | .25 | .20 | .10 | .05 | .02 | .01 |
| .50 | .26 | .17 | .12 | .10 | .05 | .02 | .01 | ... |
| .70 | .14 | .10 | .07 | .06 | .03 | .01 | ... | ... |
| .90 | .05 | .03 | .02 | .02 | .01 | ... | ... | ... |
| .95 | .02 | .02 | .01 | .01 | ... | ... | ... | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .65 | .72 | .77 | .80 | .88 | .93 | .96 | .98 |

[4] We could use the normal curve to determine $d^2$, $P_{\text{I}}$, and $P_{\text{II}}$. This method is illustrated in case (iii).

Case (iii). $\nu_1 \neq \nu_2, \qquad \sigma_1 \neq \sigma_2$.

The critical region is given by the inequality

$$\frac{p_2}{p_1} = \frac{\sigma_1}{\sigma_2} e^{\frac{1}{2}\left[\left(\frac{z-\nu_1}{\sigma_1}\right)^2 - \left(\frac{z-\nu_2}{\sigma_2}\right)^2\right]} \geqq k.$$

If we assume $\nu_2 > \nu_1$ and $\sigma_2 > \sigma_1$, this can be reduced to

$$\left(\frac{z-\nu_1}{\sigma_1}\right)^2 - \left(\frac{z-\nu_2}{\sigma_2}\right)^2 \geqq k',$$

and finally

$$\left[z - \frac{\nu_1\sigma_2^2 - \nu_2\sigma_1^2}{\sigma_2^2 - \sigma_1^2}\right]^2 \geqq d^2.$$

If we let

$$\bar{x} = \frac{\sum_{i=1}^{N_1} x_i}{N_1}, \qquad \bar{y} = \frac{\sum_{j=1}^{N_2} y_j}{N_2},$$

$$s_1^2 = \frac{\sum_{i=1}^{N_1}(x_i - \bar{x})^2}{N_1 - 1}, \qquad s_2^2 = \frac{\sum_{j=1}^{N_2}(y_j - \bar{y})^2}{N_2 - 1},$$

$$\alpha = \frac{\bar{x}s_2^2 - \bar{y}s_1^2}{s_2^2 - s_1^2},$$

we can approximate the critical region by the inequality

$$[z - \alpha]^2 \geqq d^2.$$

In order to determine $d^2$, $P_{\text{I}}$, and $P_{\text{II}}$, we have

$$P[(z-\alpha)^2 \geqq d^2 | H_1] = P[|z-\alpha| \geqq d | H_1] = \tag{7}$$

$$1 - \frac{1}{\sqrt{2\pi}} \int_{\frac{-d-(\bar{x}-\alpha)}{s_1}}^{\frac{d-(\bar{x}-\alpha)}{s_1}} e^{-t^2/2} dt = P_{\text{I}},$$

and

$$P[(z-\alpha)^2 \leqq d^2 | H_2] = P[|z-\alpha| \leqq d | H_2] = \tag{8}$$

$$\frac{1}{\sqrt{2\pi}} \int_{\frac{-d-(\bar{y}-\alpha)}{s_2}}^{\frac{d-(\bar{y}-\alpha)}{s_2}} e^{-t^2/2} dt = P_{\text{II}}.$$

The efficiency of classification associated with this case is given in table 3 below.

TABLE 3

VALUES OF $P_{\text{I}}$ AND $P_{\text{II}}$ FOR THE CRITICAL REGION $(z - \alpha)^2 \geq d^2$ CORRESPONDING TO VARYING DIFFERENCES IN MEANS AND STANDARD DEVIATIONS

$[\bar{y} = \bar{x} + hs_1, (h > 0); s_2 = \lambda s_1, (\lambda > 1)]$

$h = 1$

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | |
|---|---|---|---|---|---|
| | $\lambda = 2$ | $\lambda = 3$ | $\lambda = 4$ | $\lambda = 5$ | $\lambda = 10$ |
| .01 | .72 | .55 | .46 | .38 | .20 |
| .05 | .61 | .48 | .38 | .31 | .16 |
| .10 | .51 | .40 | .30 | .25 | .13 |
| .20 | .41 | .31 | .24 | .20 | .08 |
| .50 | .23 | .16 | .13 | .09 | .05 |
| .70 | .12 | .10 | .07 | .06 | .03 |
| .90 | .06 | .03 | .02 | .02 | .01 |
| .95 | .02 | .01 | .01 | .01 | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .67 | .73 | .77 | .80 | .88 |

$h = 3$

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | |
|---|---|---|---|---|---|
| | $\lambda = 2$ | $\lambda = 3$ | $\lambda = 4$ | $\lambda = 5$ | $\lambda = 10$ |
| .01 | .37 | .40 | .37 | .33 | .19 |
| .05 | .25 | .30 | .29 | .26 | .15 |
| .10 | .19 | .25 | .24 | .22 | .12 |
| .20 | .14 | .19 | .18 | .17 | .08 |
| .50 | .06 | .10 | .10 | .09 | .05 |
| .70 | .03 | .06 | .06 | .05 | .03 |
| .90 | .01 | .02 | .02 | .02 | .01 |
| .95 | .01 | .01 | .01 | .01 | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .84 | .81 | .81 | .82 | .89 |

$h = 2$

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | |
|---|---|---|---|---|---|
| | $\lambda = 2$ | $\lambda = 3$ | $\lambda = 4$ | $\lambda = 5$ | $\lambda = 10$ |
| .01 | .56 | .52 | .43 | .36 | .19 |
| .05 | .43 | .39 | .34 | .29 | .16 |
| .10 | .36 | .34 | .29 | .24 | .13 |
| .20 | .27 | .26 | .22 | .19 | .08 |
| .50 | .14 | .14 | .12 | .10 | .05 |
| .70 | .08 | .08 | .07 | .05 | .03 |
| .90 | .03 | .03 | .02 | .02 | .01 |
| .95 | .01 | .01 | .01 | .01 | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .76 | .76 | .79 | .81 | .88 |

$h = 4$

| $P_{\text{I}}$ | $P_{\text{II}}$ | | | | |
|---|---|---|---|---|---|
| | $\lambda = 2$ | $\lambda = 3$ | $\lambda = 4$ | $\lambda = 5$ | $\lambda = 10$ |
| .01 | .20 | .29 | .31 | .29 | .19 |
| .05 | .12 | .19 | .23 | .22 | .14 |
| .10 | .09 | .17 | .19 | .19 | .12 |
| .20 | .06 | .13 | .15 | .15 | .08 |
| .50 | .02 | .07 | .08 | .07 | .05 |
| .70 | .01 | .04 | .05 | .04 | .03 |
| .90 | ... | .01 | .02 | .01 | .01 |
| .95 | ... | .01 | .01 | .01 | ... |
| $P = 1 - P_{\text{I}}$<br>$= 1 - P_{\text{II}}$ | .91 | .85 | .83 | .83 | .89 |

## 4. Comments and conclusions

A study of the efficiency tables given in section 3 leads one to the following observations:

*a*) Within the range of values most frequently encountered, a unit increase in either mean differences or standard deviation differences (the unit being taken as the smaller of the two standard deviations) results in about a 5 per cent increase in efficiency in classification.

*b*) The larger the difference between standard deviations the less effective the difference in means. In fact, if the standard deviation difference is extremely large, the effect of the mean difference virtually disappears.

*c*) For a constant difference in means there is a standard deviation difference which minimizes the efficiency of classification.

Although the techniques of classification presented are but approximate methods and hence need further study and refinement, only two considerations will be mentioned here:

*a*) In the event $N_1$ and $N_2$ are not large, what statistic should one use in classification, and what is the probability distribution of this statistic? It may happen that some modification of the exact statistic would yield satisfactory results and greatly simplify the distribution problem.

*b*) A method of computing the index of efficiency $P$ directly needs to be developed for the various distributions encountered in classification problems.