CHAPTER 8

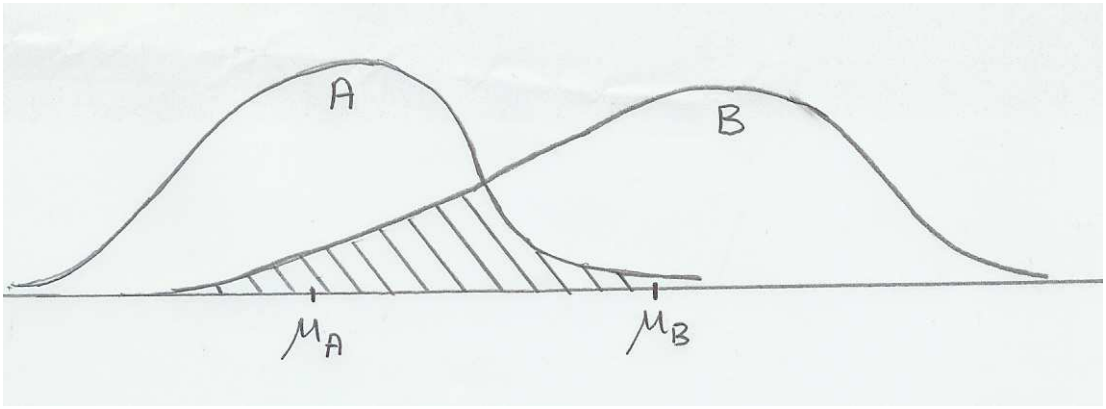
Effect size and Group overlapping

8.1 Introduction

When one wants to compare two populations with respect to a continuous response variable (as, for example, IQ, diastolic blood pressure, test marks of a person), the standardized difference δ is an appropriate effect size index (see Chapter 4). The quantity ρ^2_{pb}, i.e., the proportion variance attributed to population membership (see Chapter 5), can also be used as an index. In the case where more than two populations are compared, one can use δ for contrasts and η^2 as a generalization of ρ^2_{pb} as effect size-indices (see Chapter 6). All of these indices are based op the assumption of homogeneity of variances of the populations and, with the exception of the indices $\Delta_1, \Delta_2, \Delta_m, \delta_g$ and δ_c (see paragraph 4.3), there are no other effect size-indices available when the variances of the populations are heterogeneous. The same problem also exists when one looks at multivariate populations: all the indices (as discussed in Chapter 7) assume that the populations have the same covariance matrices.

A possible solution would be to obtain an index which is based on population overlapping. In Figure 8.1 the shaded area is the overlapping between the two population distributions A and B. Note that the populations need not necessarily be normally distributed with equal variances. Clearly the overlapping is inversely proportional to the difference in location of the two distributions (for example $\mu_B - \mu_A$). This means that the *non-overlapping* of the population distributions is large then μ_A and μ_B differ greatly, and it is small if the population means do not differ very much.

Figure 8.1: Overlapping van two population distributions



In this chapter we will discuss how the overlapping between populations can be converted into an effect size index. The case involving homogenous variances for two and more populations and with one and more variables will be explored. However, since effect size indices already exist for these cases, we will attempt to determine the relationship between the new indices and the existing indices. Thereafter the index in the cases of unequal variances or dissimilar covariance matrices will be discussed. As a result, we will first need to look at the classification of observations and the definition of a hit rate.

8.2 **Distance and classification**

Suppose that population g has a p-variable mean vector or *centroid* $\boldsymbol{\mu}_g = \left(\mu_{1g}, \mu_{2g}, ..., \mu_{pg}\right)$ and covariance matrix $\boldsymbol{\Sigma}_g$. The so called *Mahalanobis distance* of a vector of observations, $\boldsymbol{x}_u = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_s\right)$, belonging to an object u (e.g., a person) from the centroid of g can then be written as

$$\Delta_{ug} = \left[\left(\boldsymbol{x}_u - \boldsymbol{\mu}_g\right)\boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{x}_u - \boldsymbol{\mu}_g\right)'\right]^{1/2} \tag{8.1}$$

In the univariate case with p = 1, this reduces to

$$\Delta_{ug} = \frac{x_u - \mu_g}{\sigma_g},$$

where μ_g and σ_g are the mean and SD of population g.

Now, to classify $\boldsymbol{x_u}$ as belonging to one of the k populations, we make use of predictive discriminant analysis (or PDA). According to Huberty (1994:45) the purpose of PDA is as follows:

Suppose that we draw random samples from k populations of sizes $n_g, g = 1,...,k,$ which are made up of measurements on each of the $N\left(=\sum_g n_g\right)$ objects. By using this $N \times p$ data matrix, we want to determine from which one of the k populations it is the most likely to draw the (N + 1)-th object.
To determine the population from which a forthcoming object, with observed value $\boldsymbol{x_u}$, is drawn, it is assumed that the populations each have a multivariate normal distribution. With the aid of these assumptions it is now possible to make use of maximum-likelihood methods. In Huberty (1994: chapter IV) the background of this method is discussed in detail. However, for the purposes of this manual the discussion in the following paragraph will be sufficient.

8.2.1 Prior probabilities

Let π_g denote the proportion of objects in the k populations which come from population g. Thus, if an object is randomly chosen from all the populations, then π_g is the probability that it came from population g. This probability is called prior or “a priori” because it is known beforehand, i.e., before any samples are drawn.

If the k populations’ sizes are not known, π_g can be obtained in two ways:

(a) Choose it according to good judgement and past experience: the researcher knows from experience that, for example, objects from population 1 are twice as common those from population 2. In this case he/she would choose $\pi_1 = 2/3$ and $\pi_2 = 1/3$.

(b) Make the assumption that the samples' sizes are proportional to the population sizes, then we can use $\pi_g = p_g = n_g / N$.

(c) Choose $\pi_g = 1/k$, i.e., equal for all k populations.

8.2.2 Equal population covariance matrices

If we assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, the distance in (8.1) becomes Δ^*_{ug}, where $\boldsymbol{\Sigma}_g$ is replaced with $\boldsymbol{\Sigma}$. This distance is estimated by:

$$D^*_{ug} = \left[\left(\boldsymbol{x_u} - \bar{\boldsymbol{x}}_{\boldsymbol{g}} \right) \boldsymbol{S}^{-1} \left(\boldsymbol{x_u} - \bar{\boldsymbol{x}}_{\boldsymbol{g}} \right)' \right]^{1/2} \tag{8.3}$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ and **S** are the sample centroid and pooled sample covariance matrix respectively. By using maximum likelihood methods, the following classification rule is obtained (Huberty, 1994: 61-62):

Assign object u to population g if

$$D^{*2}_{ug} - 2\ell n(p_g) < D^{*2}_{ug'} - 2\ell n(p_{g'}), \tag{8.4}$$

for all $g \neq g'$.

This is known as the *linear classification rule* .

8.2.3 Unequal population-covariance matrices

Here Δ_{ug} is estimated by:

$$D_{ug} = \left[\left(\boldsymbol{x_u} - \boldsymbol{x_g} \right) \boldsymbol{S}_{\boldsymbol{g}}^{-1} \left(\boldsymbol{x_u} - \bar{\boldsymbol{x}}_{\boldsymbol{g}} \right)' \right]^{1/2}, \tag{8.5}$$

where $\boldsymbol{S}_g$ is the sample covariance matrix of population g. Maximum likelihood methods in this case produces the *quadratic classification rule*:

Assign object u to population g if:

$$\ell n\left|S_g\right|+D_{ug}^2-2\ell n(p_g)<\ell n\left|S_{g'}\right|+D_{g'}^2-2\ell n(p_{g'}), \tag{8.6}$$

for all $g \neq g'$.

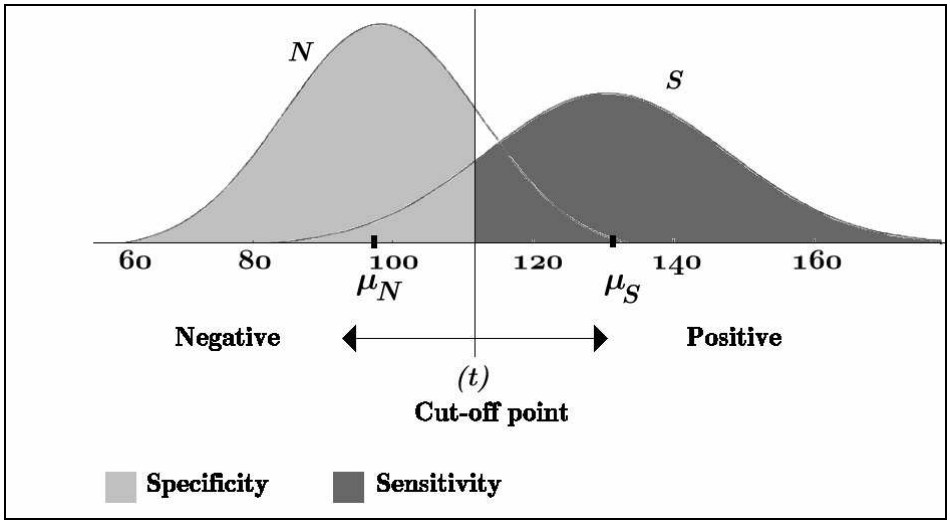
8.2.4 Two univariate populations: classification with ROC-analysis

To classify objects in this case, methods for receiver-operating characteristic (ROC) curves can be utilised.

Suppose that a disease or abnormality is studied and individuals are categorized as “positive” if they exhibit the disease or abnormality and “negative” otherwise. Other examples include clinical psychologists that would like to classify individuals as depressive or normal, or a bank that want to classify clients asking for loans as being potentially risky or not. These classifications are typically made through the use of “gold standard” diagnostic tests that are possibly expensive and/or time-consuming. If a method (also called a screening test) existed to identify the diseased/abnormal/risky individuals, that was simpler and cheaper than the “gold standard”, then it would be important to know how trustworthy it was as a predictor for diseased/abnormal/risky individuals. In further discussion we will refer to the population of sick (S) and the non-sick (N) individuals for the individuals that have a disease or not, have an abnormality or not, or are considered risky or not. Typically the measurements made for these screening tests produce a continuous value (i.e., a value that varies over a certain interval) and a cut-off or threshold value is often used to classify individuals, e.g., values above the threshold value indicate the presence of a disease, while values below it indicate the absence of the disease. Figure 1

provides a graphical representation of the distributions of populations S and N's screening test measurements.

Figure 8.2



If individuals are classified according to their actual status (according to some golden standard) as well by the screening test, then it produces the following 2x2 – frequency table:

	Actual status		
Screening test	Sick (S)	Not-sick (N)	Total
+	A (true pos.)	B (false pos.)	A + B (test +)
-	C (false neg.)	D (true neg.)	C + D (test -)
Total	A + C (sick)	B + D (not-sick)	N=A+B+C+D

In this table there are A sick individuals that reacted positively to the screening test. If it is expressed as a proportion of all the sick individuals $(A+C)$, then it gives the *sensitivity*, i.e.,

$$\text{Sensitivity} = \frac{A}{A+C}, \tag{8.7}$$

the proportion of correctly classified positives (i.e., where sick individuals test positively).

Similarly, there are D individuals correctly classified as not-sick; when this is expressed as a proportion with respect to the total not-sick individuals it is called the *specificity*, i.e.,

$$\text{Specificity} = \frac{D}{B+D}, \tag{8.8}$$

the proportion of correctly classified negatives (i.e., where not-sick individuals test negatively).

A good screening test should have a high sensitivity as well a high specificity, because the opposite would be unfavourable. That is, to classify a sick person as not-sick (a total of C individuals) is unfavourable; similarly it is also unfavourable to classify a person as being not-sick if they are sick (a total of B individuals).

In the populations (as shown in Figure 8.2) the area under the S-distribution (sick individuals) to the right of the cut-off point denotes the *sensitivity* and the area under the N-distribution to the left of the cut-off point is called the *specificity*. Ideally the two distributions would be completely separated and the cut-off point chosen such that both the sensitivity and specificity are equal to 1.

Choice of the optimal cut-off point:

The ROC-curve shows, for a sequence of cut-off values (t), the relationship between the proportion of true positives (tp) versus the proportion

of false positives (fp). The question now is whether or not there is an optimal value for t? One method is to use the Youden index YI:

$$\begin{aligned} \mathrm{YI} &= \max_t (tp - fp) \\ &= \max_t (tp + tn - 1) \ , \end{aligned} \tag{8.9}$$

i.e., the maximum value of the sum of the sensitivity (tp) and specificity (tn) minus 1. This index is a descriptive measure of the ROC-curve. The *optimal value* of the *cut-off point* t *is thus obtained when the sum* $tp + tn$ *is at its maximum.*

For a given t, the distribution functions for X_N and X_S are:

$F(t) = \mathrm{P}(\mathrm{X}_N \le t) = tn$ and $G(t) = \mathrm{P}(\mathrm{X}_S \le t) = 1 - tp$,

therefore it follows that

$$\begin{aligned} \mathrm{YI} &= \max_t (tp + tn - 1) \\ &= \max_t \left(F(t) - G(t) \right) . \end{aligned} \tag{8.10}$$

$\mathrm{YI} > 0$ implies that $F(t) \ge G(t)$ for each t, which means that the distribution of X_S lies largely to the right of the distribution of X_N (see, for example, Figure 8.2). If $\mathrm{YI} \le 0$ it means that the screening test is no better than simply randomly classifying individuals as positive or negative.

The estimated optimal t can thus be found where the estimated difference $\hat{F}(t) - \hat{G}(t)$ is a maximum. The following four methods for determining the optimal t (denoted by t^*) are discussed by Krzanowski & Hand (2009), paragraph 9.4:

(a) binormal method:

When F and G are both normal

$$\mathrm{YI} = \max_{\mathrm{t}} \left[\Phi\left(\frac{\mathrm{t} - \mu_{\mathrm{N}}}{\sigma_{\mathrm{N}}} \right) - \Phi\left(\frac{\mathrm{t} - \mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}} \right) \right],$$

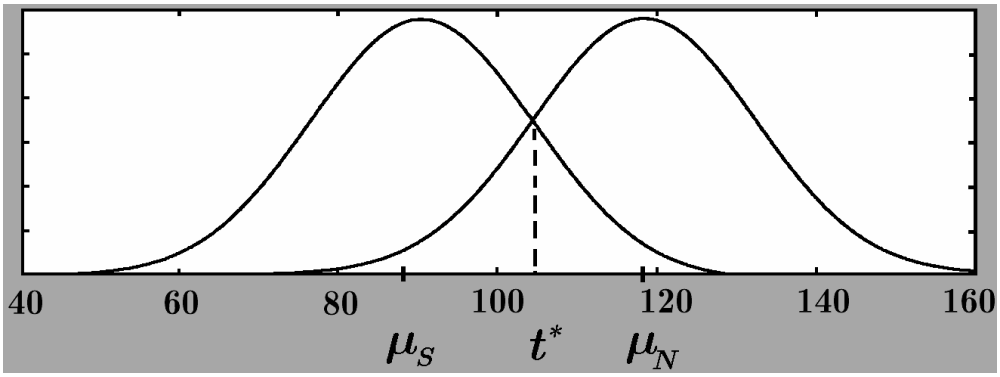
which, after setting the first derivative to zero and solving, we get:

$$t^* = \frac{\mu_S\sigma_N^2 - \mu_N\sigma_S^2 - \sigma_N\sigma_S\sqrt{(\mu_N-\mu_S)^2 + (\sigma_N^2-\sigma_S^2)\ln(\sigma_N^2/\sigma_S^2)}}{(\sigma_N^2-\sigma_S^2)} \quad . \tag{8.11}$$

If $\sigma_N^2 = \sigma_S^2 = \sigma^2$, then $t^* = ½(\mu_N + \mu_S)$, (8.12)

i.e., the optimal value lies halfway between the means of the distributions and where the normal density functions intersect each other (see Figure 8.3).

Figure 8.3:



To estimate t^* with $\hat{t}^*$, the estimators $\hat{\mu}_N$, $\hat{\mu}_S$, $\hat{\sigma}_N^2$ and $\hat{\sigma}_S^2$ are substituted in (8.11).

(b) Transformed normal method:

The assumption that $G(t)$ and $F(t)$ are normally distributed is sometimes unrealistic which means that, like in the estimation of the ROC-curve, an appropriate monotone transformation (Box-Cox transformation) can be applied to X to achieve normality. Just as the ROC-curve is invariant under monotone transformations, YI is also invariant. The optimal value t^* can then be determined as in (a) above, but on the distribution of Y , after which it can be back-transformed in terms of X .

(c) Empirical method:

F and G can be estimated with their empirical distribution functions

$$\hat{F}(t) = n'_{N(t)} / n_N$$

$$\hat{G}(t) = n'_{S(t)} / n_S \ , \tag{8.13}$$

where $n'_{A(t)}$ is the number of individuals from population A such that its X values are smaller than or equal to t.

The value $\hat{t}^*$ is then the t-value in a sequence of values that makes $\hat{F}(t) - \hat{G}(t)$ a maximum.

(d) Kernel estimation methods:

Here $F(t)$ and $G(t)$ are determined using kernel estimators of the density functions f_s and f_g (see the nonparametric estimation of ROC-curves discussed above).

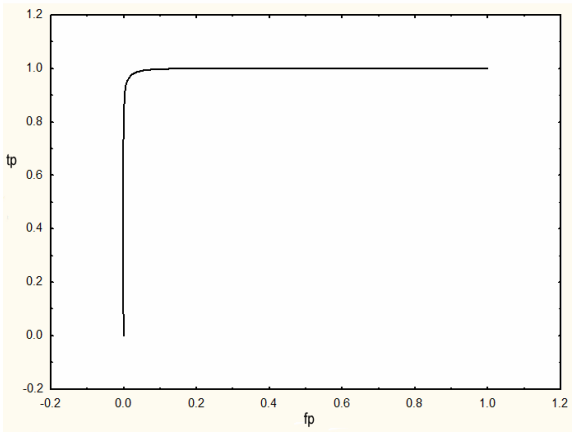
From the ROC-analysis the two following measures can be used as effect sizes: (a) the Youden-index, and (b) the area under the ROC-curve.

- Youden-index: From its definition in (8.10) it is clear that YI's value can vary between 0 and 1, with value 0 when the distributions of the two populations are identical, and value 1 when there is no overlap whatsoever.

- Area Under the ROC-curve (AUC): Consider Figure 8.4:

Figure 8.4

(a)



(b)

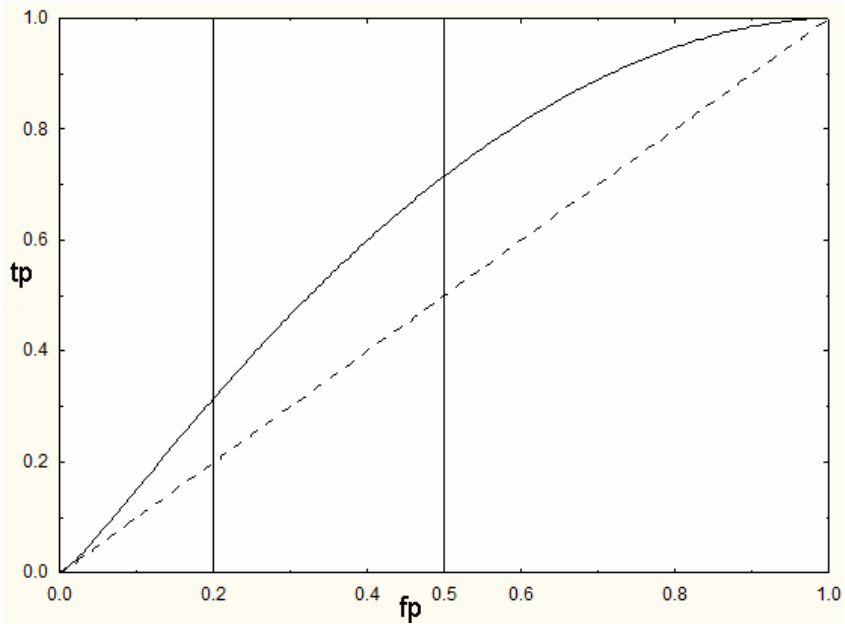


Figure 8.4(a) is obtained if the populations in Figure 8.2 are both normally distributed, population S has mean and standard deviation $\mu_S = 4, \sigma_S = 1,$ and population N has mean and standard deviation $\mu_N = 0$ and $\sigma_N = 1$. Here the populations are completely separated since the density function of the S population lies almost entirely to the right of the density function of the N

population. This ROC curve illustrates the near best possible curve obtainable; the optimal cut-off point between the two distributions can easily be chosen. The other extreme is illustrated by the diagonal line (dashed line) in Figure 8.4(b) where it is used to indicate the hypothetical situation where both distributions are assumed to be $N\ (0;1)$ and thus indistinguishable from one another.

In this case individuals from each population are equally likely to be classified as positive or negative.

Figure 8.4(b) also illustrates a ROC-curve (solid line) that could be obtained from a situation similar to Figure 1; an appropriate optimal cut-off point can then be chosen from this graph.

If one considers Figure 8.4, it is clear that the area under the ROC-curve in (a) is approximately 1, in (b) it is between 0,5 and 1 in the case of the solid line, and exactly 0,5 in the case of the dashed line. This "Area Under the Curve" is denoted by AUC and is used as *a measure of the ability to discriminate between the distributions* S and N. Larger values of AUC indicate a greater discriminatory ability. The AUC value 0,5 indicates that the one is unable to distinguish between S and N.

AUC can also be interpreted as follows:

Suppose that an individual is randomly chosen from each of the populations S and N and the screening scores are X_S and X_N, then:

$$\text{AUC} = \text{P}(X_S > X_N)\ ,$$

which means that the AUC is the probability that X_S is larger than X_N. In terms of Figure 8.4 this probability is close to 1 if the two populations are easily distinguishable (Figure 8.4(a)), whereas the probability is 0,5 if the population distributions completely overlap (Figure 8.4(b)). The AUC above is also known as the Gini-index.

Further discussions of how AUC can be estimated from ROC-analysis, etc., are found in the document "Using ROC-analysis to determine correct on

continuous variables", which can be downloaded from the following web address: www.nwu.ac.za/af/p-statcs/index.html .

8.3 **Hit rate and its estimation**

A *hit rate* is the proportion of correctly classified objects in all the populations.

Huberty (1994: Chapter VI) distinguish between three sorts of hit rates:

(a) *Optimal* hit rate $P^{(o)}$: This is the hit rate if the classification rule is based on the known population centroids and covariance matrices (i.e., $\boldsymbol{u_g}$ and Σ_g are known).

(b) *Actual* hit rate $P^{(a)}$: The expected hit rate of an upcoming sample (or test sample) where the classification rule is based on the training sample. It is also known as the conditional hit rate.

(c) *Expected* hit rate $P^{(e)}$: The expected proportion of correct classifications in all possible samples of size $N = \Sigma_g n_g$. Now we have $P^{(e)} = E\left(P^{(a)}\right)$. This hit rate is also called the *unconditional* hit rate and is of interest before any samples are drawn.

We will now look at the estimation of the hit rate in different cases.

8.3.1 Two univariate normal populations with homogeneous variances

Cohen uses the measure U_2 as the proportion of population B which is larger than the same proportion of population A (the proportion of the shaded area relative to B's total area in Figure 8.1). With the effect size $\delta = \left|\mu_B - \mu_A\right| / \sigma$ in

terms of the two populations' means and common SD, the distributions can be seen as normal $N(0;1)$ and $N(\delta;1)$ without loss of generality. This means that

$$U_2 = \Phi(\delta/2), \tag{8.14}$$

where $\Phi(x)$ is the cumulative distribution function of a $N(0;1)$- distribution.

Example 8.1

Consider Example 4.2 where $\mu_B = 111$, $\mu_A = 105$, $\sigma = \sigma_A = \sigma_B = 10$ are the mean IQ's and SD's of populations A and B respectively. With $\delta = |\mu_B - \mu_A| / \sigma = |111 - 105| / 10 = 0{,}6$, it follows that

$$U_2 = \Phi\left(\frac{0{,}6}{2}\right) = 0{,}618 \quad ,$$

which means that a proportion of 0,618 of population B has larger IQ's than the same proportion of A (see Figure 8.2 again).

Note:

Cohen (1969, 1977, 1988)'s Table 2.2.1 provides, for selected values of δ, values for U_2. For the guideline values of $\delta = 0{,}2$ $0{,}5$ and $0{,}8$ for, small, medium and large effects, the corresponding values for U_2 are $U_2 = 0{,}54$, $0{,}60$ and $0{,}66$.

According to Huberty & Holmes (1983) U_2 (which they call P_c) is the probability of a correct classification and can be estimated by:

$$P_c = \Phi(\hat{\delta}/2), \tag{8.15}$$

where $\hat{\delta} = |\bar{x}_A - \bar{x}_B| / s$, is the sample-effect size (see Chapter 4).

To maximize P_c, the following classification rule can be used for $\bar{x}_A < \bar{x}_B$ *:*

Assign object u to population A, if $x_u < \frac{1}{2}(\bar{x}_A + \bar{x}_B)$ otherwise assign it to B.

Example 8.2

For Example 4.3 the samples drawn from A and B had means 11 and 13 with variances 5 and 7,5, while the sample size was 5. Under the assumption of homogeneity of variances, we got $\delta = 0,8$ and thus $P_c = \Phi\left(\hat{\delta}/2\right) = \Phi(0,4) = 0,66$.

The classification rule is then:

Classify person u in population A if $x_u < \frac{1}{2}(11+13) = 12$, otherwise classify it in B.

If the variances are not assumed to be homogeneous and sample variances and means are substituted in (8.11), the optimum cut off point is

$$t^* = \frac{13\times 5 - 11\times 7,5 - \sqrt{5\times 7,5}\sqrt{(13-11)^2 + (5-7,5)\ln(5/7,5)}}{(5-7,5)} = 12,48,$$

which differs from the mean of 11 and 13.

□

8.3.2 Two multivariate normal populations with equal covariance-matrices

Huberty (1994: 83-86) generalize Cohen's U_2 with the optimum hit rates for population A and B as:

$$P_A^{(O)} = 1 - \Phi\left(\frac{\Gamma - \frac{1}{2}\Delta^2}{\Delta}\right) \text{ and } P_B^{(O)} = 1 - \Phi\left(\frac{-\Gamma - \frac{1}{2}\Delta^2}{\Delta}\right) \qquad (8.16)$$

where $\Gamma = \ell n(\pi_B / \pi_A)$, π_g are the a priori probabilities of membership to g and Δ is the Mahalanobis distance defined as

$$\Delta^2 = (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)'. \qquad (8.17)$$

The values Γ and Δ are estimated from two random samples with $K = \ell n\left(p_B / p_A\right)$ and $\hat{D}$ where

$$\hat{D}^2 = \frac{n-m-3}{n-2} D^2 - \frac{mn}{n_A n_B} \ , \tag{8.18}$$

with m the number of variables and

$$D^2 = \left(\bar{\boldsymbol{x}}_{\boldsymbol{A}} - \bar{\boldsymbol{x}}_{\boldsymbol{B}}\right) \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{x}}_{\boldsymbol{A}} - \bar{\boldsymbol{x}}_{\boldsymbol{B}}\right)' \quad , \quad n = n_A + n_B . \tag{8.19}$$

The quantity D^2 can easily be calculated using equation (7.6) in Chapter 7 (it is denoted there by $\hat{D}^2$) using the sample version of Wilk's Λ. The quantity $\hat{D}^2$ in equation (8.18) is then equivalent to (7.7).

Then it follows that:

$$\hat{P}_A^{(O)} = 1 - \boldsymbol{\Phi}\left(\frac{K - \frac{1}{2}\hat{D}^2}{\hat{D}}\right) \ , \ \hat{P}_B^{(O)} = 1 - \boldsymbol{\Phi}\left(\frac{-K - \frac{1}{2}\hat{D}^2}{\hat{D}}\right) \tag{8.20}$$

From this expression the total population-hit rate follows

$$\hat{P}^{(O)} = p_A \hat{P}_A^{(O)} + p_B \hat{P}_B^{(O)} . \tag{8.21}$$

For the special case where $p_A = p_B$, then (8.21) reduces to:

$$\hat{P}^{(O)} = \hat{P}_A^{(O)} + \hat{P}_B^{(O)} = \boldsymbol{\Phi}\left(\hat{D} / 2\right) \ , \tag{8.22}$$

which is a generalization of (8.15).

Example 8.3

Consider Example 7.1: In this example the estimated value of $\hat{D}$ is used to compare the means of the experimental and control groups' BDI before test, after test and follow-up test scores: $\hat{D} = 4{,}42$. However, since the two groups are chosen to be equally large, the assumption $p_E = p_K$ can be made, so that

$$\hat{P}^{(O)} = \hat{P}_E^{(O)} = \hat{P}_K^{(O)} = \boldsymbol{\Phi}\left(\hat{D} / 2\right) = \boldsymbol{\Phi}\left(2{,}21\right)$$

$$= 0{,}986 \ .$$

It is almost certain that an individual can be correctly classified in the two groups if the BDI's before, after and follow-up test scores are used.

Assign object u to population g if:

$$D_{ug}^{*2} < D_{ug'}^{*2} \text{ for } g \neq g = 1,2.$$

Here the assumption of equal covariance matrices is made and

$$\bar{\boldsymbol{x}}_1' = \bar{\boldsymbol{x}}_{\boldsymbol{E}}' = (13{,}04 \;\; 8{,}72 \;\; 6{,}72) \quad , \quad \bar{\boldsymbol{x}}_2' = \bar{\boldsymbol{x}}_{\boldsymbol{K}}' = (11{,}56 \;\; 15{,}56 \;\; 16{,}36)$$

and

$$\boldsymbol{S} = \begin{pmatrix} 35{,}76 & 12{,}57 & 14{,}97 \\ 12{,}57 & 58{,}37 & 48{,}90 \\ 14{,}97 & 48{,}90 & 80{,}17 \end{pmatrix}$$ is the pooled covariance matrix.

The linear classification rule applied to this data reduces to the following rule: classify as belonging to group E if:

-4,153 + 0,178Before + 0,89After + 0,17Followup < -3,239 + 0,346Before + 0.113After -0.061Followup,

and otherwise classify it in group K. Each person's before, after and follow-up scores of BDI are substituted into the left and right in the above formula and if the left hand side is smaller than the right hand side, the person is classified in group E, otherwise it is classified in group K.

We find thus that persons 1 and 9 from group E are incorrectly assigned into group K (thus 23 out of 25 people are correctly classified), while persons 10, 14-16, 18-21 from group K are incorrectly assigned to group E (17 out of 25 correctly classified). □

8.3.3 More than two multivariate populations

In practice we usually have many more than two multivariate populations wherein objects must be classified. Under the assumption of normal populations, the classification rule in paragraph 8.2 is used. Unfortunately, the problem is that

estimators like $\hat{P}^{(O)}$ in (8.21) do not exist for the hit rate. According to Huberty (1994) the PDA can be used in two ways: internally and externally.

(a) *Internal analysis* means that the classification rule is based on the same data on which it is applied and thus objects are reclassified. The proportion of objects which are correctly classified is an estimator of the hit rate and is called the *apparent or resubstituted hit rate*. This hit rate is biased for any of $P^{(o)}$, $P^{(a)}$ or $P^{(e)}$ and *overestimates* the hit rate. This method is used by many well-known statistical computer packages such as Statistica, SPSS, BMDP and SAS. The output is thus easily obtained but, for the reasons outlined above, care should be taken when interpreting the estimated hit rate.

(b) *External analysis* can be split into the so called hold-out method, the leave-one-out (abbreviated L-O-O) method and the maximum-posteriori-probability method.

(i) The hold-out method involves removing or holding out a test sample by randomly choosing it from the original sample. The remaining data (called the training sample) is used to construct the classification rule with which the elements within the test sample are classified to determine the hit rate. This method only provides good estimates of $P^{(a)}$ in cases where the test sample is the same size as the training sample. The abovementioned computer packages can all be used to calculate these hit rates.

(ii) The leave-one-out (L-O-O) method involves leaving out one object and then the linear classification rule is based on the remaining $N-1$ objects' observations. The object is then classified in to one of the g populations. This procedure is repeated for each of the N objects over all the populations and the hit rate is then determined afterwards. This hit rate is,

strictly speaking, not an estimator for any of $P^{(o)}$, $P^{(a)}$ or $P^{(e)}$, because it based on a sample size of $N-1$ instead of a sample of size N. However, when N is sufficiently large it can be used as an estimator for $P^{(a)}$. The SAS procedure called DISCRIM can be used to determine this hit rate by invoking the CROSSVALIDATE option. SPSS, on the other hand, can be used with the 'Leave one out' option. Unfortunately, STATISTICA does not have any options for the leave-one-out method.

(iii) The Maximum-posterior-probability method (abbreviated M-P-P method) is yet another alternative method for estimating $P^{(a)}$. It is calculated as the mean of all the objects' maximum posterior probabilities:

$$\hat{P}^{(a)} = \frac{1}{N}\sum_{u=1}^{N} max\left\{\hat{P}\left(1|\boldsymbol{x_u}\right), \hat{P}\left(2|\boldsymbol{x_u}\right), \ldots, \hat{P}\left(k\,|\boldsymbol{x_u}\right)\right\}, \qquad (8.23)$$

where $\hat{P}\left(g\,|\boldsymbol{x_u}\right)$ is the estimated posterior probability that object u falls in population g. This probability can be estimated via an internal or external analysis. Both SAS's DISCRIM procedure and SPSS's DISCRIMINANT can determine (with the aid of the internal methods) the values of $\hat{P}\left(g\,|\boldsymbol{x_u}\right)$. The estimator $\hat{P}^{(a)}$ is then called the M-P-P/I estimator. SAS's DISCRIM procedure, along with the CROSSVALIDATE option, estimates $\hat{P}\left(g\,|\boldsymbol{x_u}\right)$ by employing the external L-O-O-method and the estimator $\hat{P}^{(a)}$ is then known as M-P-P/L-O-O. According to Huberty (1994) this method is preferable if one can assume multivariate normality. If these assumptions are in doubt then the simple L-O-O method would be preferable.

Table 8.2 gives the different classifications by using the above mentioned methods for the data in Chapter 3, Example *F* with the 3 activity groups as populations from which samples of $n_1 = 694$, $n_2 = 227$ and $n_3 = 441$ $(N = 1362)$ were drawn.

8.4 **Effect size index for correct classification**

The hit rate, estimated using one of the methods described in the previous paragraph, gives us an index which can be used to judge the success of correct classification over all the populations. To judge this hit rate, it is first necessary to compare it with the so called chance classification's probability. This is the

probability of incidental classification when one does not use any data at all, also known as the *chance-hit rate*. According to Huberty (1994) there are usually two ways of determining the chance-hit rate, namely,

(a) the proportional chance-criterion and

(b) the maximum chance-criterion.

These two methods will now be discussed.

Table 8.2: Classifications with different methods:

		Classification in groups			
(a) **Internal / Linear**		**1**	**2**	**3**	**Total**
	1	549 (78,1)	1 (0,1)	151 (21,8)	694
Groups	**2**	96 (42,3)	0 (0,0)	131 (57,7)	227
	3	169 (38,3)	0 (0,0)	272 (61,7)	441
	Total	807 (59,3)	1 (0,1)	554 (40,7)	1 362
	% Error	21,9	100,0	38,3	40,2
(b) **L-O-O / Linear**					
	1	540 (77,8)	1 (0,1)	153 (22,1)	
Groups	**2**	96 (42,3)	0 (0,0)	131 (57,7)	
	3	169 (38,3)	0 (0,0)	272 (61,7)	
	Total	805 (59,1)	1 (0,1)	556 (40,8)	
	% Error	22,2	100,0	38,3	40,4
(c) **Internal / Non-linear**					
	1	207 (29,8)	66 (9,5)	421 (60,7)	
Groups	**2**	25 (11,0)	20 (8,8)	182 (80,2)	
	3	27 (6,1)	14 (3,2)	400 (90,7)	
	Total	259 (19,0)	100 (7,3)	1 003 (73,7)	
	% Error	70,2	91,2	9,3	54,0
(d) **L-O-O / Non-linear**					
	1	206 (29,7)	66 (9,5)	422 (60,8)	
	2	26 (11,5)	13 (5,7)	188 (82,8)	
	3	30 (6,8)	19 (4,3)	392 (88,9)	
	Total	262 (19,2)	98 (7,2)	1 002 (73,6)	
	% Error	70,3	94,3	11,1	55,1
	A priori-probability	0,51	0,17	0,32	

8.4.1 Proportional chance-criterion

When we have k populations of the same size from which samples of equal size n are drawn, the chance-hit rate is clearly $1/k$ for each population and the expected frequency of a hit per population is $n = n\,1/k + n\,1/k + ... + n\,1/k$. In the general case of populations with unequal sizes, equal chance hit rates of $1/k$ are replace with p_g, i.e., the estimated a priori probability of membership to population g. When we also have unequal sample sizes, i.e., n_g, $g = 1,...,k$, then the expected frequency of a hit for population g is $e_g = p_g n_g$, so that the chance-frequency of a hit over all the populations is then

$$e = \sum_{g=1}^{k} e_g = \sum_{g=1}^{k} p_g n_g \,. \qquad (8.24)$$

Let $N = n_1 + n_2 + ... + n_k$, then the chance-hit rate over all the populations is:

$$H_e = \frac{e}{N} = \frac{1}{N} \sum_{g=1}^{k} p_g n_g \qquad (8.25)$$

Example 8.4

In Chapter 3, Example *F,* the three activity groups form the populations from which the samples of size $n_1 = 694$, $n_2 = 227$ and $n_3 = 441$ $(N = 1362)$ are drawn. Suppose that the researcher chooses the *a priori* probabilities to be *0,5, 0,25* and *0,25*. Now, the chance-hit frequencies, are *0,5* x *694 = 347* for activity group *1, 0,25* x *227 = 56,75* and *0,25* x *441 = 110,25* for groups 2 and 3 respectively. Thus, the total chance hit frequencies are *347 + 56,75 + 110,25 = 514*, so that the chance-hit rate for all the groups is $H_e = 514 / 1362 = 0,377$. If we assume

that the *a priori* probabilities are proportional to the sample sizes, then it follows that the total chance-frequency of a hit becomes:

$$\left(\frac{694}{1362}\times 694\right)+\left(\frac{227}{1362}\times 227\right)+\left(\frac{441}{1362}\times 441\right)$$
$$= 353{,}6 + 37{,}8 + 142{,}8$$
$$= 534{,}2,$$

and that the total chance-hit rate is now $H_e = 534{,}2 / 1362 = 0{,}392$.

This means if the *1362* men are randomly split into groups, i.e., without using any data, then we could expect a hit rate of 38-40%. □

8.4.2 Maximum chance criterion

In this situation the chance-hit rate H_e is simply taken to be equal to the maximum of the different estimated *a priori* probabilities:

$$H_e = max\left(p_1, p_2, ...p_k\right) \qquad (8.26)$$

According to Huberty & Lowman (2000) this criterion is most applicable with 2 groups when the a priori-probabilities are radically different.

Example 8.5

In Example 8.4 we calculated $H_e = 0{,}5$ using the pre-specified a priori-probabilities. Now, the proportional a priori-probabilities are *694 / 1362 = 0,51 ; 227 / 1362 = 0,17 and 441 / 1362 = 0,32*, and we get $H_e = max\ \left(0{,}51\ ;\ 0{,}17\ ;\ 0{,}32\right) = 0{,}51$. □

8.4.3 Statistical testing of the frequency of a hit

If the frequency of a hit of population g in the classification frequency table is denoted by n_{gg}, then the total frequency of a hit is given by:

$$o = \sum_{g=1}^{k} n_{gg} \quad .$$

Under the null hypothesis of incidental classification, it follows that

$$z = \frac{o - e}{\sqrt{e(N-e)/N}} \sim N(0;1).$$

The testing of this null hypothesis produces the *p*-value: $p = P(Z \geq z)$, where Z follows a $N(0;1)$ distribution.

The lower $(1-\alpha)100\%$ confidence bound for the actual frequency of a hit is thus (Huberty, 1994: 105):

$$o\text{-}z_{\alpha}\sqrt{e(N-e)/N}, \tag{8.27}$$

where z_{α} is the $(1-\alpha)$-th percentile of a $N(0;1)$ distribution.

For population g it follows similarly that

$$o_g - z_{\alpha}\sqrt{e_g(n_g - e_g)/n_g} \tag{8.28}$$

which is the lower $(1-\alpha)100\%$ confidence bound for the actual frequency of a hit.

Example 8.6:

In Example 8.4 the proportional chance criterion was $e = 534,2$ while from Table 8.1 (b) (where the L-O-O / Linear method was used) it follows that o = *540 + 0 + 272 = 812* .

$$z = \frac{812 - 534,2}{\sqrt{534,2(1362-534,2)/1362}} = \frac{277,8}{\sqrt{324,67}} = 15,42$$

so that $p < 0,0001$. The 95% confidence-lower bound for the actual frequency of a hit is then:

$$812 - 1{,}645\sqrt{534{,}2\left(1362 - 534{,}2\right)/1362}$$
$$= 812\text{-}1{,}645\sqrt{324{,}67}$$
$$= 812 - 29{,}6 = 782{,}4$$

Applying the chosen a priori-probability for population 3, we get $e_3 = 0{,}25 \times n_3 = 0{,}25 \times 441 = 110{,}25$. Then, from Table 8.1(b) the observed frequency of a hit is $o_3 = 272$, we find that $p < 0{,}0001$ and the 99% confidence lower bound for the actual frequency of a hit for population 3 is:

$$272 - 2{,}33\sqrt{110{,}25\left(441\text{-}110{,}25\right)/441}$$
$$= 272 - 2{,}33\sqrt{82{,}69}$$
$$= 272 - 21{,}19 = 250{,}8.$$

This means that the total frequency of a hit can be as low as *782,4* with *95%* probability, while the frequency of a hit for population 3 can be as small as *250,8* with *95%* probability. □

8.5 **Effect size index: Better-than-chance**

By comparing the actual or observed hit rate H_o with the chance hit rate H_e, the value H_o is adjusted for incidental correct classification of objects. The size of the chance error rate $1 - H_e$, in comparison to the observed error rate $1 - H_o$, the following effect size index is obtained (see Huberty & Lowman, 2000, Huberty, 1994):

$$I = \frac{\left(1\text{-}H_e\right) - \left(1\text{-}H_o\right)}{1\text{-}H_e} = \frac{H_o - H_e}{1 - H_e}. \qquad (8.29)$$

From the definition of the effect size index, *I* can also be described as an index for the proportional reduction in error of the “better-than-chance” index.

Note that the index depends on the definition of "chance" as reflected by H_e, which is, in turn, dependent on the estimation of the a priori probabilities.

Example 8.7:

For Example 8.4's classification of persons within the 3 populations, Table 8.1 displays four different observed error rates $(1\text{-}H_o)$. We use the chance error rates based on the proportional a priori-probabilities, $1-H_e = 1-0,392 = 0,608$, and the following values for *I* can be obtained:

Methods	*p*	$1\text{-}H_o$	$1\text{-}H_e$	*I*
(a) Internal/Linear	<0,0001	0,402	0,608	0,339
(b) L-O-O/Linear	<0,0001	0,404	0,608	0,336
(c) Internal/Non-linear	<0,0001	0,540	0,608	0,112
(d) L-O-O/Non-linear	<0,0001	0,551	0,608	0,094

□

It would appear from Example 8.7 that methods (a) and (b) provide better classifications based on higher values of *I*. Note that in all cases $p < 0,0001$, which means that the null hypothesis of incidental classification is rejected throughout for large samples. This indicates that the classifications are not the results of simple coincidence, however, this is not necessarily a good thing. In order to judge the success of the classifications, the index *I* can be used. The value *0,34* obtained from method (a) in Example 8.7 implies that there is a 34% reduction in the error rate when using this method of classification in comparison with a plain chance classification.

8.6 Relationship between proportion variance (η^2) and the better-than-chance index (*I*)

8.6.1 Homogenous variances or covariance matrices

To try and get an intuitive "feel" for the index *I*, Huberty & Lowman (2000) considered six of the thirteen dependent variables. They compared groups 1 and 2 for each of the variables for the BISBEY-data in Huberty (1994). The variances are assumed to be homogenous throughout, so that it reduces to the univariate two group case with homogenous variances.

By using *0,333* and *0,667* as a priori probabilities and choosing to employ the maximum-chance-criterion, we obtain: $H_e = 0{,}667$. The *t*-values for the *t*-test on each of the variables, provided in Table 8.3 (Huberty & Lowman's Table 1), and the hit probabilities H_o are both used in the linear classification rule. The proportion variance η^2 is estimated by using equation (5.24) in Chapter 5.

Table 8.3: Results of univariate 2-group comparisons with homogenous variances

t	p	$\hat{\eta}^2$	H_o	I
-1,07	0,286	0,010	0,698	0,09
-1,46	0,147	0,018	0,681	0,04
-3,58	0,001	0,101	0,707	0,12
-3,74	0,000	0,109	0,698	0,09
-6,29	0,000	0,258	0,810	0,43
-8,12	0,000	0,366	0,836	0,51

The Pearson correlation (r) between $\hat{\eta}^2$ and I is found to be *0,90,* while the Spearman rank correlation (r_s) of *0,81* indicates a strong monotone relationship.

In the same manner Huberty & Lowman (2000) proceeds by using the same data to determine the relationships between the estimated η^2 values and the I-index in the following cases:

(a) Univariate, 3-group comparisons with homogenous variances: *r = 0,97*
(b) Multivariate, 2-group comparisons with homogenous covariance-matrices: *r = 0,95*
(c) Multivariate, 3-group comparisons with homogenous covariance-matrices: *r = 0,98*

While these results do not necessarily hold in general, they do provide a good indication of whether or not a positive linear relationship exists between η^2 and I in the case of homogenous variances and covariance matrices.

8.6.2 Heterogeneous variances of covariance matrices

The estimated proportion variance η^2, which, in the previous paragraph, was related to I as effect size-indices, can no longer be used because homogenous variances of covariance matrices must be assumed. Huberty & Lowman (2000) therefore attempted to correlate I with test statistics which are used to test the null hypothesis of equal means when heterogeneous variances or covariance matrices are assumed. For a chosen data set they tried to ascertain this correlation in the following cases:

(a) Univariate, 2-group comparisons with heterogeneous variances: Test statistic *J* (James's 2nd order test) with I (the non-linear / L-O-O-method with maximum-chance criterion) produces the correlation *r = 0,89*.
(b) Univariate, 3-group comparisons with heterogeneous variance: *r = 0,88*
(c) Multivariate (4 variables), 2-group comparisons with heterogeneous covariance matrices Test statistic *T* (from Yao) with I:

r = 0,97

(d) Multivariate 3 variables), 3-group comparisons with heterogeneous covariance matrices: Test statistic *S* (from Johansen) with *I*: *r = 0,84*

8.7 **Guideline values for the index *I***

When one compares two populations, then Huberty & Holmes (1983) agree with Cohen (1969) that the standardized difference of $\delta = 0,2$ indicates a small effect (see paragraph 4.5, Chapter 4). According to their table 2, the expected hit rate $P^{(e)}$ is approximately *0,55.* If we assume equal a priori-probabilities for the **two populations**, it follows that $H_e = 0,5$, while if the estimated value of $P^{(e)}$ using H_o is *0,55,* then we find that $I = \frac{0,55 - 0,5}{0,5} = 0,1$. Huberty & Holmes feel that a medium effect for classification with an expected hit rate of $0,65$ (i.e., $I = 0,3$) should correspond, which is equivalent to $\delta = 1,0$. Further, they require that $P^{(e)} = 0,75$ is a large effect (i.e., $I = 0,5$), where $\delta = 1,5$.

Huberty & Lowman (2000) suggest the following guidelines, based on exploratory analyses for univariate 2-group classification with homogenous variances:

Effect	
Small	$I < 0,1$
Medium	$0,15 < I < 0,25$
Large	$I > 0,3$

In the *k*-group case, Huberty & Lowman (2000) suggest using these same guidelines. They also provide guidelines for the other cases. Table 8.4 is a summary of these guidelines:

Table 8.4: Guidelines for better-than-chance index

		Number	Effect		
Univariate		**Populations**	**Small**	**Medium**	**Large**
	Homogenous	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
Variances		k	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
	Heterogeneous	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
		3	$I < 0,05$	$0,10 < I < 0,20$	$I > 0,25$
Multivariate					
	Homogenous	2	$I < 0,15$	$0,2 < I < 0,3$	$I > 0,35$
Covariance-		3	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
Matrices	Heterogeneous	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
		3	$I < 0,05$	$0,10 < I < 0,20$	$I > 0,25$

In summary, Huberty & Lowman (2000)'s suggestion that, for all cases, $I \leq 0,1$ should be considered a small effect and $I \geq 0,35$ should be considered a large effect. However, they warn that their suggestions are based on a restricted number of data sets and that they did not investigate cases with more than *3* to *4* populations.

8.8 **Uses of the index** I

The effect size index I can be used in ways. First, it can be used as an index for the success of the classification rule in the discriminant analysis. This usage would be employed when the primary interest is in determining whether the classification rule is able to correctly classify future observations. A second usage of I is to use it as an effect size index instead of using η^2 (and other special cases, e.g., δ) when heterogeneous variances occur. While η^2 can be

influenced by the number of groups and the number of observations *(N*), I 's value is never directly influenced by *N* (Huberty & Lowman, 2000).

When we have heterogeneous variances, the indices $\Delta_1, \Delta_2, \Delta_m, \delta_g$ and δ_c can be used in the univariate, 2-group case. For any other situation however, (see Chapters 6 and 7 of this manual), we assume homogeneity of variances. Here I can be effectively used when it is based on the quadratic classification rule.

We usually assume normality when performing statistical tests on the mean vector and is followed up by the estimation of the effect size η^2. Essentially, the assumption of normality is not necessary in the use of classification rules, but usually the rules result from it. Huberty (1994: Chapter X) provides some methods for performing discriminant analysis when normality is not assumed. If the variables are continuous, but are non-normal, we can use, for example, rank transformations and nearest neighbours analyses. These methods are available in SAS (among other computer packages).

For categorical, nominal variables there are, according to Huberty, two possibilities. First, one can create c - 1 dummy variables (with values *0* and *1*) for each of these types of variables, where *c* is the number of categories in the variable. The problem with this method is that the final number of dummy variables can cause the analysis to become intractable. In these cases ordinary classification rules can be used. A second possibility is to perform a Fisher-Lancaster analysis (Huberty, 1994: 153) where each variable category represents a score. This transformed data can once again be dealt with using ordinary classification rules.

The abovementioned methods cover almost all situations (heterogeneity of variance and non-normality among other) making the better-than-chance index I a more general index than η^2 for example.

More research is necessary to determine the relationship between I and η^2 in a larger variety of situations than discussed in Huberty and Lowman (2000). Also, guideline values for I are still very tentative and should be made clearer so that they can be made more general.

A SAS program which can be used to calculate I (*Groepoorvleueling.sas*) is available on the web page of this manual.