Sex determination of the feral house cat *Felis catus* using multivariate statistical analyses

S.H.C. du Toit, R.J. van Aarde and A.G.W. Steyn
University of Pretoria

An analysis of skull measurements of feral domestic cats *Felis catus* by principal component and discriminant function analyses resulted in the formulation of a function discriminating with a high degree of accuracy (89,5%) between the sexes. The function is defined as u = 2,366 (basal length) + 2,427 (palatilar length) − 0,759 (zygomatic width) + 1,336 (cranium width) − 29,984. The statistical methods employed are described in a simple and informative way.

S. Afr. J. Wildl. Res. 1980, 10: 82 – 87

Die analisering van skedelmates van die wilde huiskat *Felis catus* deur die gebruik van hoofkomponent- en diskriminantfunksie analise het aanleiding gegee tot die formulering van 'n funksie wat met 'n hoë mate van akkuraatheid (89,5%) onderskei tussen die geslagte. Dié funksie word gedefinieer as u = 2,366 (basale lengte) + 2,427 (palatinum lengte) − 0,759 (zygomatiese wydte) + 1,336 (kranium hoogte) − 29,984. Die statistiese metodes wat gebruik is, word eenvoudig beskryf.

S.-Afr. Tydskr. Natuurnav. 1980, 10: 82 – 87

S.H.C. du Toit
Department of Statistics, University of Pretoria, Pretoria 0002
R.J. van Aarde*
Mammal Research Institute, University of Pretoria, Pretoria 0002
A.G.W. Steyn
Department of Statistics, University of Pretoria, Pretoria 0002
*To whom all correspondence should be addressed

Received 15 September 1979; accepted 20 March 1980

Attempts to analyse sex specific population dynamic parameters, making use of information obtained from skulls collected in the field, is often hampered by the inability to determine the sex of such material. This problem arose in a study on the population ecology of the feral domestic cat, *Felis catus*, inhabiting Marion Island. In the present paper the techniques of principal component analysis (PCA) and discriminant function analysis (DA) are used to analyse measurements taken from skulls in an attempt to determine the sex of the collected material.

This technique, described briefly to introduce the novice to this field, can also be used in other studies with similar data. However, due to the fact that the use of these multivariate statistical methods are often not very clear to biologists lacking an in depth knowledge of statistical tools, the steps, in an abbreviated form, are included in an Appendix. The statistical equations included here are kept to a minimum but are nevertheless essential to an adequate understanding of the techniques applied and the subsequent analysis of the results obtained.

Both principal component and discriminant function analyses have been used as statistical tools for well over half a century. Biological application has become quite numerous since the widespread use of the electronic computer, and presently several multivariate data analysis texts dealing with these are available. Examples are Cooley and Lohnes (1971), Kshirsagar (1972), Morrison (1976) and Tatsuoka (1971). Various computer programmes have been developed performing the necessary calculations entailed in these two methods. A popular one (the one used in this paper) is the Statistical Package for the Social Sciences or as commonly kown, SPSS (Nie, Hull, Jenkins, Steinbrenner and Bent 1975). Great care should be exercized in interpreting the PCA and DA computer output, since inadequate explanations regarding the actual output and its meaning accompany the SPSS text. Based upon their personal experience, the authors present explanations, and draw conclusions from the graphical and numerical SPSS output.

Material and methods

During the course of an 18 month study at Marion Island (46°54′S, 37°45′E) situated in the south Indian Ocean, 230 intact adult cat skulls (skulls with permanent dentition) were collected. Of these, 124 skulls (80♂♂ and 44♀♀) were obtained from culled animals, and 106 were

Appendix 1 Numbers of eagles destroyed according to the bounty system in certain divisions of the Cape Province between 1931 and 1946 (excluding 1933). From provincial notices (Cape), C.P.A. archives.

	DATE																	
DIVISION	1931	1932	–	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	TOTAL
Barkly West										3	1	3			1		8
Bedford	5													9	7		21
Heidelberg					6	5	4	11		1	6	1	3	10	6	8	61
Humansdorp					20	13			20	20	19	36	38	24		36	226
Jansenville	3										14	12	12	10	1	21	73
Lady Grey												8			12		20
Namaqualand													16	16			32
Pearston				26	25		21	48	56	24	39	44	78	33	52	31	477
Prince Albert																2	2
Riversdale	8	16				7	8	13	9	2	3	4	11	6	3	6	96
Somerset East	35	25															60
Steytlerville					1	5	16	19	13	6	19	36	15	19	37	31	217
Swellendam	2	6			7	9		7	2	3	6	3	1	3	2	1	52
Tarkastad		1		2	5	2	5		6	11	9	9	5	10	10	29	104
TOTAL	53	48	?	28	64	41	54	98	106	70	116	156	179	140	131	165	1 449

Appendix 2 Numbers of eagles destroyed according to the bounty system in certain divisions of the Cape Province between 1947 and 1955. From provincial notices (Cape), C.P.A. archives. 1 = 'Eagles'; 2 = 'Berghane'; 3 = Martial Eagles/Lammervangers. For Tawny Eagles/Kouvoëls see foot note 1.

YEAR	1947			1948			1949			1950			1951			1952			1953			1954			1955			
DIVISION	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	TOTAL
Beaufort West		26			41	1		20	1		19	2		15			13	7		12	2		14			8		180
Bedford	81			106			116					67	40			52			54			–	–	–	–	–	–	516
Caledon		2		–	–	–	2			–	–	–		1	1		2		–	–	–	–	–	–	–	–	–	8
Calvinia		15	3		6	9		12	25		14	12		11	20		11	89		5	70		6	38		1	3	350
Carnarvon	(2)-	–	–			3			4		1						1		–	–	–		2		–	–	–	11
Ceres	4			5			2			7			5			5			3				2			10		43
Cradock			6			8			11			12			30	6					25			9			17	124
De Aar			2			4		2		5					1	–	–	–	–	–	–	–	–	–	–	–	–	14
Graaff-Reinet	30			31			47			17			21			14			16			26			1	–	–	203
Iay		7			20			19			19			16			12			20			13			13		139
Heidelberg			10			7			11			4			7	4			4			9					5	61
Herbert			2			1			1			6			3			6			10			1			6	36
Humansdorp	32					26			33			30			29			20			12			14			21	217
Jansenville			11	12			14			35					7			7			3			89			4	182
Maraisburg			2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2
Mossel Bay			1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1
Pearston			35			22			25			37	27			29			22			12					1	210
Philipstown			7			7			8			7			6			5	–	–	–			6			1	47
Piketberg	–	–	–		8			8			8			13		–	–	–	–	–	–	–	–	–	–	–	–	37
Prince Albert		3			9			2			7			2			3			16			4			5		51
Riversdale			6			1			9			3			3			2	–	–	–			5	–	–	–	29
Somerset East	24			22			33			24			20			32			17			22			37			231
Steynsburg			2			1			4			3	–	–	–	–	–	–			1			1	–	–	–	12
Steytlerville			36			15			15			10			13			4			4			10		5		122
Stutterheim		1			10			22			12			20		1	1		1	2	2	–	–	–	–	–	–	72
Sutherland		22			14			11			4			13			12			10			10			9		105
Swellendam	3			4			–	–	–	4			4			2			2			–	–	–	2			21
Tarkastad	4			12			8			8			10			8			3			4			–	–	–	57
Uniondale	–	–	–	–	–	–	–	–	–	–	–	–			3		5			3				8			18	37
Willowmore		20	9		31	16		35	18		31			2		–	–	–	–	–	–	–	–	–	–	–	–	162
Wodehouse			1			1			2			3			3			2			3	–	–	–	–	–	–	15
TOTAL	178	96	133	192	139	122	222	131	167	100	115	196	127	93	126	153	60	142	122	68	132	73	51	181	40	50	76	3 285

1. Tawny Eagle/Kouvoël returns were tendered as follows: Carnarvon — 1948(1), 1950(1), 1951(5) and Steytlerville — 1950(5).
2. Not clear whether no returns submitted or whether returns submitted but no eagles destroyed.

Table 1 A matrix of correlation coefficients for ten skull measurements of 80 male adult cat skulls from which principal components were obtained.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,000									
X_2	0,872	1,000								
X_3	0,916	0,882	1,000							
X_4	0,762	0,720	0,802	1,000						
X_5	0,765	0,670	0,734	0,543	1,000					
X_6	0,534	0,471	0,448	0,398	0,574	1,000				
X_7	0,273	0,246	0,232	0,207	0,313	0,239	1,000			
X_8	0,464	0,451	0,393	0,316	0,513	0,483	0,644	1,000		
X_9	0,512	0,478	0,510	0,410	0,683	0,449	0,325	0,395	1,000	
X_{10}	0,552	0,496	0,461	0,493	0,327	0,233	0,043	0,182	0,199	1,000

unsexed skulls found in the field. These skulls were all that were available during the period of research. The spatial distribution of the collected skulls and the living animals were approximately the same. In the case of the skulls collected death was presumed to be from natural causes, and they therefore provided a sample resulting from natural mortality.

All skulls were measured, the measurements made being based on those defined by De Blase and Martin (1974). These measurements were: basal length (X_1), condylobasal length (X_2), basilar length (X_3), palatilar length (X_4), zygomatic width (X_5), mastoidal width (X_6), interorbital constriction (X_7), cranium width (X_8), cranium height (X_9) and maxilar toothrow length (X_{10}). By performing principal component and discriminant function analyses on data obtained from the skulls of animals of which the sex was known, a function discriminating between the sexes was obtained. The measurements from skulls collected in the field, were substituted in the function, and the sex of these skulls thereby determined.

Results and discussion

Principal component analysis

In the SPSS programme FACTOR used during this analysis the principal components or factors are obtained from the matrix of correlation coefficients. For illustrative purposes this matrix, as has been obtained for measurements made on male skulls, is given in Table 1.

Table 2 Sample variances (eigen-value) and cumulative relative sample variances (as percentage) of the principal components based on the correlation matrix for males.

Factor	Eigen-value	Cumulative percentage of the variation accounted for
1	5,552	55,5
2	1,434	69,9
3	0,819	78,1
4	0,630	84,4
5	0,549	89,8
6	0,364	93,5
7	0,264	96,1
8	0,219	98,3
9	0,105	99,4
10	0,064	100,0

Table 2 shows the sample variance and relative sample variances, (the last as a percentage) accounted for by each factor. For further analysis only the first two factors are considered because this programme uses the arbitrary rule of considering only the factors with corresponding sample variances larger than one. However it is evident that 70% of the variance in the data system is accounted for by these two factors. Therefore, without loss of too much information the original 10 dimensional data system can be transformed to a two dimensional space.

Table 3 Correlations between the original variables (X_1 to X_{10}) and the first two Principal Components (PC). Standardized coefficients are shown in the last two columns.

Variables	Factor 1 PC No 1	Factor 2 PC No 2	Standardized Coefficient PC No 1	PC No 2
X_1	0,934	−0,195	0,396	−0,163
X_2	0,887	−0,204	0,377	0,171
X_3	0,904	−0,247	0,384	−0,206
X_4	0,795	−0,303	0,337	−0,254
X_5	0,850	0,107	0,361	0,089
X_6	0,649	0,222	0,275	0,185
X_7	0,420	0,716	0,178	0,598
X_8	0,620	0,602	0,263	0,503
X_9	0,669	0,245	0,284	0,205
X_{10}	0,540	−0,452	0,229	−0,378

Table 3 shows the standardized coefficents in equations (1) and (2) (cf. Appendix) and also the sample correlation between the original variables and the first two principal components (equation (3) cf. Appendix). The absolute values of these standardized coefficients give an indication of the importance of each variable in the corresponding factor. It is of interest to note that while the first component can be considered as a general summarizing factor indicating the relative importance of the individual variables in an overall study, the second component compares X_1, X_2, X_3, X_4 and X_{10} (length parameters) with X_5, X_6, X_7, X_8, and X_{10} (width parameters).

The classification of variables into two groups (or clusters) according to the sign of their correlations with factor 2 are represented in Fig. 1, indicating that groups A and B represent width and length variables respectively. For comparative purposes the corresponding two dimensional plot for females has been superimposed on that of males (Roman symbols represent the female correlation). The information in Fig. 1 furthermore illustrates that if an investigator would wish to restrict a study to a fewer number of variables, a subset might be chosen from each of the clusters A and B.

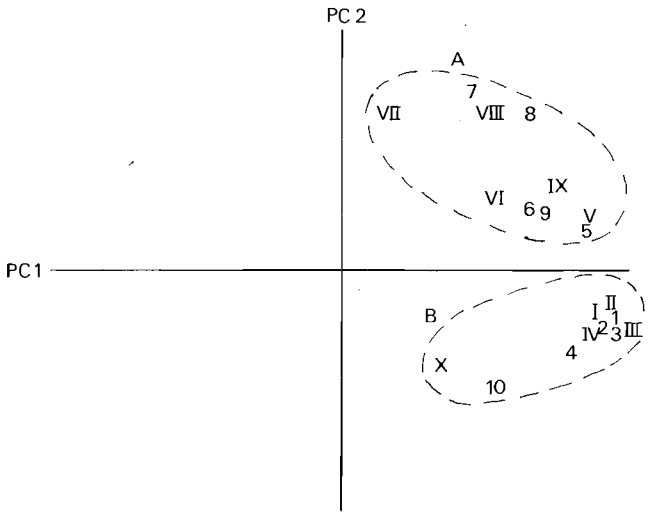

Fig. 1 Plot of correlations between original variables and principal component (PC) 1 against correlations between original variables and principal component (PC) 2. Roman symbols represent the female correlates with the original variables X_1 to X_{10}.

Discriminant analysis

For the problem under consideration the SPSS programme DISCRIMINANT was used. Two groups (g = 2) were considered, these being females (group 1; n = 44) and males (group 2; n = 80). Means and standard deviations for the 10 variables (X_1 to X_{10}) measured are given in Table 4.

The computation of a discriminant function (DF) based on all 10 the variates measured, yield an 89,5% correct classification according to sex. However, an evaluation of the effectiveness of this function on a statistical basis was not possible since the underlying assumptions for statistical inference could not be established. A test conducted for multivariate normality of the two samples (Mardia 1974), which involved the calculation and testing of measures of multivariate skewness and kurtosis, indicated that the data could not be regarded as samples from a 10 dimensional normal population.

Further analysis indicated that approximately 10% of all observations (in 10 dimensions) could be regarded as 'outliers' (an observation considered as an 'outlier' if the value of at least one of the 10 variables differed by more than three sample standard deviations from its corresponding group sample mean).

Subsequently the test for multivariate normality was performed on the data with the 'outliers' omitted. All the multivariate measures, apart from the skewness in the male sample lay well within the acceptable limits. A test for equality of covariance matrices (Morrison 1976) for different groups suggests that the two group covariance matrices differed significantly.

Consequently it was decided to choose a subset of variables in such a way that both the assumptions of multivariate normality and equality of covariance matrices were met, subject to the condition that they are truly representative of the original variables. On thi basis X_1 and X_4 were chosen to represent the length variables while X_5 and X_8 were chosen to represent the width variables and a discriminant analysis was performed on them.

Table 4 Means and standard deviations for the 10 skull variables measured

Variables		Females (Group 1) n = 44	Males (Group 2) n = 80
Basal length	X_1	8,481 ± 0,337	9,278 ± 0,354
Condylobasal length	X_2	7,865 ± 0,301	8,600 ± 0,318
Basilar length	X_3	7,068 ± 0,279	7,747 ± 0,296
Palatilar length	X_4	3,097 ± 0,150	3,387 ± 0,133
Zygomatic width	X_5	6,263 ± 0,317	6,787 ± 0,368
Mastoidal width	X_6	3,899 ± 0,208	4,167 ± 0,146
Interorbital constriction	X_7	3,110 ± 0,177	3,194 ± 0,171
Cranium width	X_8	4,203 ± 0,126	4,535 ± 0,125
Cranium height	X_9	3,513 ± 0,134	3,700 ± 0,176
Maxilar toothrow length	X_{10}	2,646 ± 0,196	2,858 ± 0,131

To evaluate the DF we use a special case of H_{01}. If k = 0, the null hypothesis (H_{01} cf. Appendix) becomes H_{01}: The DF derived affords no discrimination between groups. The value of the statistic (11) is $\chi^2 = 101{,}71$ (df = 4). The probability of obtaining this χ^2-value under H_{01} is 0,000 (to 3 decimal places). H_{01} is therefore rejected and we conclude that the DF obtained can be used with confidence in the classification of unknown individuals. The standardized DF coefficients (equation (10) cf. Appendix) are especially important in this problem since only one DF was derived. These coefficients (0,821 for X_1; 0,337 for X_4; − 0,266 for X_5; 0,167 for X_8) indicate the relative importance of each of the four characteristics in discriminating between sexes.

If the discriminating power of any selected characteristic is in question, a F-test can be carried out. (according to equation (13) cf. Appendix). Under H_0 (cf. Appendix) the estimated F-values have the F-distribution with 1 and 122 degrees of freedom. It has therefore been concluded that all four variables have discriminating power.

Unstandardized DF coefficients used in the calculating of the group centroids (−1,545 and 0,850 for females and males respectively) were 2,366, 2,427, −0,759 and 1,336 for X_1, X_4, X_5 and X_8 respectively. Discrimination between sexes is based on these unstandardized function coefficients and a constant calculated at −29,984. To test the ability of these coefficients to discriminate between the sexes the function defined as:

$$U = 2{,}366\,X_1 + 2{,}427\,X_4 - 0{,}759\,X_5 + 1{,}336\,X_8 - 29{,}984 \quad \ldots \text{(A)}$$

were used to reclassify the 44 female and 80 male skulls according to sex with a centroid of −1,545 for females and 0,850 for males. If for example, skull measurements X_1, X_4, X_5 and X_8 are obtained from a skull of unknown sex, the U value can be obtained by use of (A). A skull is classified as that of a female if this calculated value is closest to the group 1 centroid, and as male when closest to group 2 centroid.

It is informative to note (Table 5) that 89,52% of the known cases were grouped correctly when these DF coefficients were used. This gives a measure of the confidence with which this DF can be used to classify skulls of these domestic cats according to sex, using skull parameters.

By using this discriminant function (A) the 106 skulls collected in the field could be classified according to sex. Forty-five of these skulls were classified as females and 61 as males (Table 5). This classification of the study material according to sex now enabled the worker to use this material in the analyses of sex specific population dynamic characteristics.

Table 5 Classification of skulls of known and unknown sex according to sex using the discriminant function (A).

Actual sex	Sample size	Predicted sex: Females	Predicted sex: Males
Females	44	40	4
Males	80	9	71
Unknown	106	45	61

Percentage of known cases correctly classified: 89,52%

It is of interest to note that adult sex ratios favouring males in populations inhabiting several sub-Antarctic Islands have been reported for Kerguelen (Derenne 1974, Pascal 1977), Macquarie (Jones 1977) and Marion Island (Van Aarde 1978). A tendency towards sex ratios in favour of males in samples for feral cats collected on continental areas were reported by Nilsson (1940), McMurray and Sperry (1941), Eberhard (1954) and Coman and Brunner (1972).

These sex ratios, showing a preponderance of males in the adult age class, could be due to an inadequacy in sampling technique, probably arising from a difference in behaviour of the sexes and/or from sex-specific mortality, causing a reduced survival rate of females.

Acknowledgements

We are grateful to the Department of Transport for providing logistical and financial support on advice of the South African Committee for Antarctic Research. This programme was carried out under the auspices of the Mammal Research Institute of the University of Pretoria and we wish to thank Prof. J.D. Skinner for his enthusiastic support and Dr. N. Fairall for reviewing the manuscript.

References

COMAN, B.J. & BRUNNER, H. 1972. Food habits of the feral house cat in Victoria. *J. Wildl. Mgmt.* 36: 848 – 853.

COOLEY, W.W. & LOHNES, P.R. 1971. Multivariate data analysis. Wiley, New York.

DEBLASE, A.F. & MARTIN, R.E. 1974. A manual of mammalogy. Wm.C. Brown Company, Dubuque, Iowa.

DERENNE, P.L. 1976. Notes sur la biologie du chat haret de Kerguelen. *Mammalia* 40: 532 – 595.

EBERHARD, T. 1954. Food habits of Pennsylvania house cats. *J. Wildl. Mgmt.* 18: 284 – 286.

JONES, E. 1977. Ecology of the feral cat, *Felis catus* (L), (Carnivora: Felidae) on Macquarie Island. *Aust. Wildl. Res.* 4: 249 – 262.

KSHIRSAGAR, A.M. 1972. Multivariate analysis. Dekker, New York.

MARDIA, K.V. 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya.* B36: 115 – 128.

McMURRAY, F.B. & SPERRY, C.C. 1941. Food of feral cats in Oklahoma: a progress report. *J. Mammal.* 22: 185 – 190.

MORRISON, D.F. 1976. Multivariate statistical methods. McGraw-Hill Inc., New York.

NIE, H.N., HULL, C., STEINBRENNER, KARIN & BENT, D.H. 1975. Statistical Package for the social sciences (2nd ed.) McGraw-Hill, New York.

NIE, H.N. & HULL, C. 1978. The SPSS Inc. Batch System, Release 8, *Update manual.*

NILSSON, N.N. 1940. The role of the domestic cat in relation to game birds in the Williamette Valley, Oregon. Unpubl. Masters Thesis, Oregon State University.

PASCAL, M. 1977. Contribution á l'etude de la structure, et de la dynamique de la population de chats harets de l'archipel des Kerguelen (40°50'L.S., 69°30'I.E.). Thesis submitted for Third cycle doctorate, Université Pierre et Marie Curie.

TATSUOKA, M.M. 1971. Multivariate analysis; techniques for educational and psychological research. Wiley, New York.

VAN AARDE, R.J. 1978. Reproduction and population ecology in the feral house cat *Felis catus* on Marion Island. *Carniv. Genet. Newsl.* 3: 288 – 316.

Appendix

Description of statistical methods employed

(i) Principal component analysis (PCA)

Suppose that p measurements have been made on each of N randomly selected individuals. Denote the p characteristics measured by the random variables X_1, X_2, . . ., X_p. PCA entails the transformation of these variables to a new set of random variables Y_1, Y_2, . . ., Y_p, the principal components (PC'S or factors) which possess the following desirable properties:

(a) The factors, each being a linear combination of the original variables, are uncorrelated.

(b) The first factor

$$Y_1 = a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p \quad \ldots (1)$$

accounts for the largest proportion of the variance in the multivariate system, the second factor

$$Y_2 = a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p \quad \ldots (2)$$

accounts for the second largest proportion of the variance etc.

Usually the first few factors account to a large extent for the variation in the data, the remaining factors contributing negligible additional information. If for example 80% of the variance in a data system of 10 responses could be accounted for by (1), it would appear that this factor provides sufficient information to confine the analysis to a one rather than to a 10 dimensional space. In this case the coefficients a_{11}, a_{21}, . . ., $a_{10,1}$ of the linear combination (1) would indicate the relative importance of each original variable in the factor Y_1.

It should be noted that when the PC's are extracted from the sample covariance matrix the coefficients in (1) and (2) are called unstandardized coefficients. If however, the PC's are extracted from the correlation matrix, they are called standardized coefficients. In the SPSS program FACTOR used during the present investigation, factors were obtained from the matrix of the correlation coefficients.

The correlation between the random variable X_i and the jth factor Y_j is given by

$$r_{X_i,Y_j} = a_{ij}\sqrt{\ell_j}/S_i,\ i = 1,2, \ldots, p, j = 1, 2, \ldots, p; \quad (3)$$

where S_i and $\sqrt{\ell_j}$ are respectively the sample standard deviation of X_i and Y_j. A plot of the points $(r_{X_i,Y_1}; r_{X_i,Y_2})$; i = 1, 2, . . ., p, enables the investigator to classify the original variables X_1, X_2, . . ., X_p into homogenous groups. Principal component analysis therefore usually serves as a first step in gaining more insight into the underlying structure which has produced a specific data set.

(ii) Discriminant Analysis (DA)

Frequently it is possible to subdivide a population into g non-overlapping groups P_1, P_2, . . ., P_g (e.g. A specific species is subdivided according to sex into g = 2 groups, these being male and female respectively.) DA can be used as a satisfactory method whereby an individual can be allocated to a specific group if (i) p characteristics have been measured on N individuals from known groups, (ii) the same measurements have been performed on the individual in question.

If a random sample of size N_k is drawn from population P_k, then we denote the p × p matrix of sums of squares and cross products of deviations from the means computed from this sample by A_k. A typical element of this matrix $(A_k)_{i,i}$, is the sum of squares of deviations from the mean of the ith characteristic measurements from group P_k; the element $(A_k)_{i,j}$ (i ≠ j) being the sum of cross products of deviations from the ith and jth means. The matrix of sums of squares and cross products 'within groups' is defined as

$$A_w = A_1 + A_2 + \ldots + A_g. \quad \ldots (4)$$

To calculate the p × p matrix A_t of total sum of squares and cross products of deviations from the grand means, the random sample of N sets of p observations from the g different groups are pooled where

$$N = N_1 + N_2 + \ldots + N_g. \quad \ldots (5)$$

The grand means are simply the means of the pooled observations and a typical element of the 'Total' matrix, $(A_t)_{i,i}$ is the sum of squares of deviations from the mean of the ith characteristic measurements using the poole observations while $(A_t)_{i,j}$ is the sum of cross products of deviations from the ith and jth means of the pooled observations.

Similar to the analysis of variance (ANOVA) the total sum of squares and cross products matrix A_t can be partitioned into two additive components:

$$A_t = A_a + A_w, \quad \ldots (6)$$

where A_a is the matrix of sum of squares and cross products of deviations of group means from the grand means, called the 'Among-groups' matrix.

Let us consider the following linear combination which will be called a discriminant function (DF)

$$\begin{aligned} U &= V_1(X_1 - \overline{X}_1) + V_2(X_2 - \overline{X}_2) \\ &\quad + \ldots + V_p(X_p - \overline{X}_p) \quad \ldots (7) \\ &= V_1X_1 + V_2X_2 + \ldots + V_pX_p + \text{constant}, \end{aligned}$$

where $\overline{X}_i$ is the grand mean of the ith characteristic. It can be shown that in order to maximize the discrimination between the g groups it is necessary to maximize the ratio λ of the 'Among groups' and the 'Within groups' quadratic forms (q_1 and q_2 respectively) with respect to the DF coefficients V_1, . . ., V_p, where

$$\lambda = q_1/q_2, \quad \ldots (8)$$

$$q_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} V_i(A_a)_{i,j}V_j \text{ and}$$

$$q_2 = \sum_{i=1}^{p}\sum_{j=1}^{p} V_i(A_w)_{i,j}V_j.$$

The solution to this optimization problem will yield $g^* = \min\{p, (g - 1)\}$ uncorrelated DF's with unstandardized DF coefficients, each being of the form (7). It can be shown that the first DF, that is the one corresponding to the largest value of λ will provide maximum discrimination between groups, the next one second largest discrimination etc. Let U_1, U_2, . . ., U_{g^*} denote these functions with corresponding variances Θ_1, Θ_2, . . ., Θ_{g^*} where in general the variance of U is given by

$$\Theta = q_3/(N - 1), \quad \ldots (9)$$

and $q_3 = \sum_{i=1}^{p}\sum_{j=1}^{p} V_i(A_t)_{i,j}V_j$.

Consider the kth group P_k with N_k sets of p measurements. When we replace $X_1, \ldots, X_p$ in (7) by the corresponding group means of group P_k, the value then obtained is denoted by C_{jk} and is called the kth group centroid for the jth DF; k = 1, 2, . . ., g. In DA the coefficients of the DF's are therefore chosen in such a way that maximum separation between group centroids is obtained. To classify an 'ungrouped' individual on the basis of p measurements, g* DF scores are calculated by substituting the values of these measurements (x_1, x_2, . . ., x_p) into the g* DF's obtained on the basis of all N observations. The individual is then classified as belonging to that group whose g* centroids are geometrically the nearest.

This analysis may also be employed as a method whereby a p-dimensional data set can be displayed graphically in one or two dimensions. If only two groups of subjects are involved in a particular research then g = 2 so that only one DF is computed. In this case the two group centroids are plotted on a straight line giving an indication of 'distance' between groups. When more than two groups are considered, one usually finds that the plot of the corresponding group centroids of the first two DF's are most important.

If each coefficient in (7) is divided by the standard deviation of U then $U/\sqrt{\theta}$ is called a standardized DF, that is

$$\begin{aligned} U/\sqrt{\theta} &= V_1(X_1 - \overline{X}_1)/\sqrt{\theta} + V_2(X_2 - \overline{X}_2)/\sqrt{\theta} \\ &\quad + \ldots + V_p(X_p - \overline{X}_p)/\sqrt{\theta} \\ &= (V_1S_1/\sqrt{\theta})Z_1 + (V_2S_2/\sqrt{\theta})Z_2 \\ &\quad + \ldots + (V_pS_p/\sqrt{\theta})Z_p, \qquad \ldots (10) \end{aligned}$$

where Z_i is the ith standardized measurement $(X_i - \overline{X}_i)/S_i$ with $\overline{X}_i$ and S_i the respective pooled within groups mean and standard deviation, i = 1, . . ., p. The coefficients $V_1S_1/\sqrt{\theta}$, $V_2S_2/\sqrt{\theta}$, . . ., $V_pS_p/\sqrt{\theta}$ are called standardized DF coefficients. These coefficients give an indication of the relative importance of the original variables in a specific DF.

Denote the various values of λ (in 8) associated with the DF's $U_1, U_2, \ldots, U_{g^*}$ by $\lambda_1, \lambda_2, \ldots, \lambda_{g^*}$ respectively. Using the assumption that the g samples have been drawn randomly from p-dimensional normal populations with common covariance matrices, the following hypothesis can be tested:

H_{01}: The last (g* − k) DF's afford negligible additional discrimination between groups.

The statistic for testing H_{01} is given by

$$\chi^2 = -\left(N - \frac{p+g}{2} - 1\right)\log_e \Lambda', \qquad \ldots (11)$$

which under the null hypothesis is approximately Chi-square distributed with (p − k) (g* − k) degrees of freedom. Λ' is known as Wilks Lambda and may be computed from

$$\Lambda' = \prod_{j=k+1}^{g^*} \frac{1}{1+\lambda_j}, \; k = 0, 1, \ldots, g^* - 1 \qquad \ldots (12)$$

As part of DA, the following null hypothesis is also usually tested: H_{02}: The g group population means of the ith characteristic, i = 1, . . ., p, are equal.

By introducing straightforward ANOVA principles the following test criterium can be derived

$$F = \frac{1 - \Lambda_i}{\Lambda_i}\left(\frac{N-g}{g-1}\right), \qquad \ldots (13)$$

where $\Lambda_i = (A_w)_{i,i}/(A_t)_{i,i}$; i = 1, . . ., p. When H_{02} is true, F has the F distribution with g − 1 and N − g degrees of freedom respectively.