USE OF MORPHOMETRIC CHARACTERS TO IDENTIFY NORTH AMERICAN AND EUROPEAN STOCKS OF ATLANTIC SALMON (Salmo salar L.)

JOHN GLENN LUTHER, B.Sc.(hon.)

ISBN 0-315-55011-2

Canada

USE OF MORPHOMETRIC CHARACTERS TO IDENTIFY NORTH AMERICAN AND EUROPEAN STOCKS OF ATLANTIC SALMON (*Salmo salar* L.)

by

John Glenn Luther, B.Sc. (hon.)

A practicum report submitted to the School of Graduate Studies in partial fulfillment of the requirement for the Degree of Master of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

May 29, 1989

St. John's, Newfoundland

ACKNOWLEDGEMENTS

Financial support was provided by the School of Graduate Studies in the form of a University fellowship and teaching assistantships from the Department of Mathematics and Statistics. The Science Banch of the Northwest Atlantic Fisheries Centre, Department of Fisheries and Oceans, provided statistical data used in this study.

Special thanks to my supervisor, Dr. Brajendra Sutradhar, for his guidance and support. Thanks also to Dr. David Reddin of the Department of Fisheries and Oceans for technical advice. The assistance of faculty and staff, especially Wanda Heath, in the Department of Mathematics and Statistics and staff at Computing Services, Memorial University, is appreciated. The encouragement of fellow graduate students cannot go without mention. Finally, I would especially like to thank my parents for continued support and encouragement throughout my graduate program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Based on samples of Atlantic salmon smolts from 13 geographically distinct home rivers, stocks from North America and Europe can be distinguished by morphometric character sets using discriminant analysis procedures. Character sets require morphometric measurements of total length, standard length, predorsal length, dorsal to adipose, head length, postorbital length, and left pectoral length. A quadratic discriminant analysis was determined to be the most appropriate technique to classify the salmon smolts as either European or North American in origin. The analysis of the morphometric characters provided strong statistical seperation between areas. A classification of groups yielded 99.65% correct classification between European and North American stocks.

Chapter 1

INTRODUCTION

1.1 Background of the Problem

A common problem in fisheries research is estimating the annual proportion of different stocks of fishes in a given fishery. In particular, since first assessment of the effect of the Greenland fishery for Atlantic salmon (Salmo salar L.) on homewater stocks and fisheries, scientists have been interested in the annual proportions of North American and European salmon in the exploited population off West Greenland. These estimates are then used to assess the effect of the West Greenland fishery on stocks and fisheries in home waters. In this context, for example, Ritter, et al. (1980) assessed the impact of the West Greenland salmon fishery on stocks and catches in North America. Their assessment indicated that exploitation of salmon at West Greenland was resulting in a reduced yield to all fisheries in homewaters per recruit. For every tonne of salmon caught at West Greenland, losses to homewater stocks and fisheries ranged from 0.54 to 1.28 tonnes.

There are other similar problems of interest. For example, identifying capelin stocks in Canadian Atlantic waters, distinguishing redfish species in the Northwest Atlantic, identifying Baltic stocks of Atlantic salmon, and identifying Newfoundland and Scottish stocks of Atlantic salmon, to name just a few.

For the estimation of proportions of stocks in mixed stock fisheries, some discrimination criterion is frequently used as the statistical tool. For example, Lear and Misra (1978) dealt with scales of adult Atlantic salmon collected from 18 river systems in eastern North America. They analyzed scale character variables including smolt age and ciruli counts and found that significant differences occurred in each of these variables between river systems. These differences were also found to be significantly related to latitude. They found that the numbers of circuli in each of the three growth zones (on the salmon scale) increased from north to south, while the smolt ages decreased from north to south. They demonstrated that there were highly significant differences between scale characteristics among samples of Atlantic salmon from northern Labrador to Maine. The reason for the Lear and Misra study was that commercial fisheries for Atlantic salmon in Newfoundland and Labrador exploit mixed stocks of fish originating in river systems in Newfoundland, Labrador, the Maritimes, Quebec and Maine, U.S.A.

Sharp, et al. (1978) performed a multivariate discriminant analysis on capelin using nine morphometric and eleven meristic variables. The samples came from the St. Lawrence estuary, the Gulf of St. Lawrence, the Grand Banks, and Notre Dame Bay, Newfoundland. The subsequent analysis of the meristic variables provided no evidence of discrete stocks. Such analysis of meristic variables offered little promise as a diagnostic tool in the classification of separate stocks of capelin in the Canadian Atlantic area. However, analysis of morphometric variables provided strong statistical separation between areas. Morphometric

measurements used were eye diameter, snout length, head length, body depth, snout-vent length, snout-dorsal origin, adipose fin base, pelvic-pectoral distance, and pectoral fin length. Only snout length, eye diameter, head length and body depth contributed significantly to the separation obtained.

Misra and Ni (1983) analyzed morphometric data from 100 deepwater redfish and 100 Labrador redfish. Twelve morphometric variables were measured - body weight, head length, snout length, interorbital width, preanal length, pectoral fin base, anal fin base, length of longest pelvic ray, length of longest pectoral ray, width of caudal peduncle, dorsal length of caucal peduncle, and standard length. They carried out a classification study of the beaked redfishes, in which the specimens of Labrador redfish were relatively smaller than those of deepwater redfish. In their study, they used a discriminant function with covariance. A discriminant function of several variables seperated the species effectively with seven morphometric characters identified as pertinent discriminators. They also found that a discriminant function with covariance seperated species better than one without covariance.

MacCrimmon and Claytor (1984) dealt with juvenile Atlantic salmon of seven river stocks in northern, north-central, central and southern Sweden. The purpose of their study was to identify the nature and extent of taxonomic diversity occurring among Baltic salmon in various Swedish rivers using meristic and morphometric data and to determine if these variables could be used for the identification of regional

and home river stocks by discriminant analysis. Morphometric variables used in the study were head length, upper jaw length, distance between pectoral and pelvic fins, distance between the pelvic and anal fins, gape width, head width, body width, head depth, body depth, caudal peduncle depth, pectoral fin length, pelvic fin length, and standard length. In their study, they determined whether or not meristic and morphometric variables could be used to identify regional and home river origins. Morphometric variables provided a better means of identification than meristic variables. However, while meristic variable differences between river stocks were less pronounced, they did have considerable power in discriminating regional stocks. They concluded that each of the Swedish river stocks examined may be regarded as distinct using morphometric variables.

Reddin (1986) used scale character variables to develop and test a statistical model to classify Atlantic salmon caught at West Greenland, as either North American or European in origin. Scale samples collected in 1980 from salmon caught in Europe and North America were used as learning samples to identify variables and form a database. More specifically, scale samples used as European standards were obtained from adult salmon of known European origin, in namely, Ireland, Scotland and Norway. Scale samples from the North American standard came from specimens sampled from commercial catches at Twillingate and Burgeo, Newfoundland. A stepwise discriminant analysis was used to select the best variables, and it was determined that a quadratic discriminant analysis was the most appropriate technique to classify

the salmon. A test sample of known origin, independent of the learning database used for the discriminant analysis, resulted in a very low misclassification rate.

MacCrimmon and Claytor (1986) based their paper on a pooled sample of 367 specimens of juvenile Atlantic slamon, from eight geographically distinct home rivers. These specimens of juvenile Atlantic salmon representative of each of four Newfoundland and four Scottish rivers were obtained during 1982. They were distinguished by meristic and morphometric variable sets using discriminant analysis procedures. Meristic variables were used along with morphometric measurements of standard length, pectoral and pelvic fin lengths, body depth, and gape width. Based on their data, only the morphometric discriminant function was highly accurate in identifying home river origins of the fish examined with the discriminating power increasing with increased fish size. The set of classification functions from these data provided a good seperation of pooled fish from the eight home rivers into their regional Newfoundland and Scottish origins. The classification of the eight home river stocks was also high, with only one river falling below a 75% accuracy. Their findings for juvenile fish indicated that morphometric data sets would seem to offer the best possibility for identifying the river of origins of adult Atlantic salmon in mixed-stock fisheries.

Finally, Kenchington (1986) analyzed a set of morphological data for two types of northwest Atlantic Redfishes, using multivariate techniques. He examined 15 morphometric variables including standard

length, snout to anal fin distance, body depth, caudal peduncle depth, head length, snout length, orbit height and inneroribital distance. Although species were significantly different, they could not be fully seperated using these variables. He suggested that electrophoretic techniques were needed for precise identifications. He also found that although the two types of redfishes of the Scotian shelf had significantly different body forms, they could not be clearly distinguished on the basis of these morphometric data. They were more distinct in their meristic characteristics. This study was initiated to reveal useful characters for discriminating between North American and European salmon, their annual variation and variability between stocks.

It will be shown how discriminant analysis of morphological characters can be used in discriminating a European from a North American origin salmon. The specimens for the study were caught as smolts in European rivers in 1969 and North American rivers in 1968 and 1969.

1.2 Plan of the Project

The plan of the project is as follows:

1. In order to study the distributional aspects of the data as discussed in Section 2.1 of Chapter 2, Exploratory Data Analysis techniques will be used. The Box and Cox (1964) method of shifted-power transformation will be used to normalize the data set.

2. (a) Discriminant functions will be developed to discriminate

   (i) North American and European origin salmon.

   (ii) All salmon originating from the five sampled European rivers.

(iii) All salmon originating from the six North American rivers sampled in 1968; and the eight North American rivers sampled in 1969.

(iv) All salmon originating from the five common North American rivers sampled in both 1968 and 1969.

(b) To verify classification procedures, the jackknife classification technique will be used to determine the bias inherent in basing classification decisions on that data set used to determine the classification functions.

(c) The observations will also be classified using canonical variables instead of the original discriminating variables. Thus, the first two canonical variables will be plotted to show the seperation of the g groups. These resulting classification boundary lines will be superimposed over the plot of cases to obtain a better picture of how cases are being classified.

3. Finally, the data of the g different groups will be combined to form a single data set (ie. the five sampled European rivers; the eight North American rivers sampled in 1969; etc.). A clustering technique will be computed to determine if the g groups are well seperated. That is, a discriminant analysis will be performed based on the clustering principle.

1.3 Data Collection and Description

All specimens of salmon smolts used in the study were collected from rivers in Europe and eastern North America during the months of May, June and July in the years 1968 and 1969. In Europe, samples were taken from Logan River, Sweden; River Almond, Scotland; River Boyne and River Lee, Ireland; and River Usk, Wales. In North

America, samples were taken from Enfield's Hatchery in Maine, U.S.A.; at the Curventon fish enumeration facility, Miramichi River; Beechwood Dam, Saint John River in New Brunswick; Koksoak River and Kaniapiskou River in Ungava Bay, Quebec; Indian River Spawning Channel, Salmon River, Harry's River and Salmonier River in Newfoundland; and Sand Hill River in Labrador. The location of these rivers are shown on the maps of figure 1.1a and 1.1b.

All specimens were kept frozen until examined. The seven morphometric variables measured on each specimen were:

(1) Total Length - the length of the salmon measured from the tip of the snout to the farthest tip of the caudal fin. The measurement is a straight line and is not taken over the curve of the body.

(2) Standard Length - the distance between the tip of the snout to the end of the vertebral column.

(3) Predorsal Length - the distance between the tip of the snout to the front structural base of the dorsal ray.

(4) Dorsal to Adipose - the distance between the back structural base of the dorsal ray to the front structural base of the adipose.

(5) Head Length - the distance from the tip of the snout to the most distant point on the opercular membrane.

(6) Postorbital Length - the distance from the closest point of the orbital socket to the most distant point on the opercular membrane.

(7) Left Pectoral Length - the distance between the two structural bases of the left pectoral ray.

These morphometric variables were measured to an accuracy of 0.1 millimetres except total length and standard length, which were measured to the nearest millimetre. Each of the measurements are shown in the diagram of figure 1.2.

Figure 1.1a Locations of North American Rivers

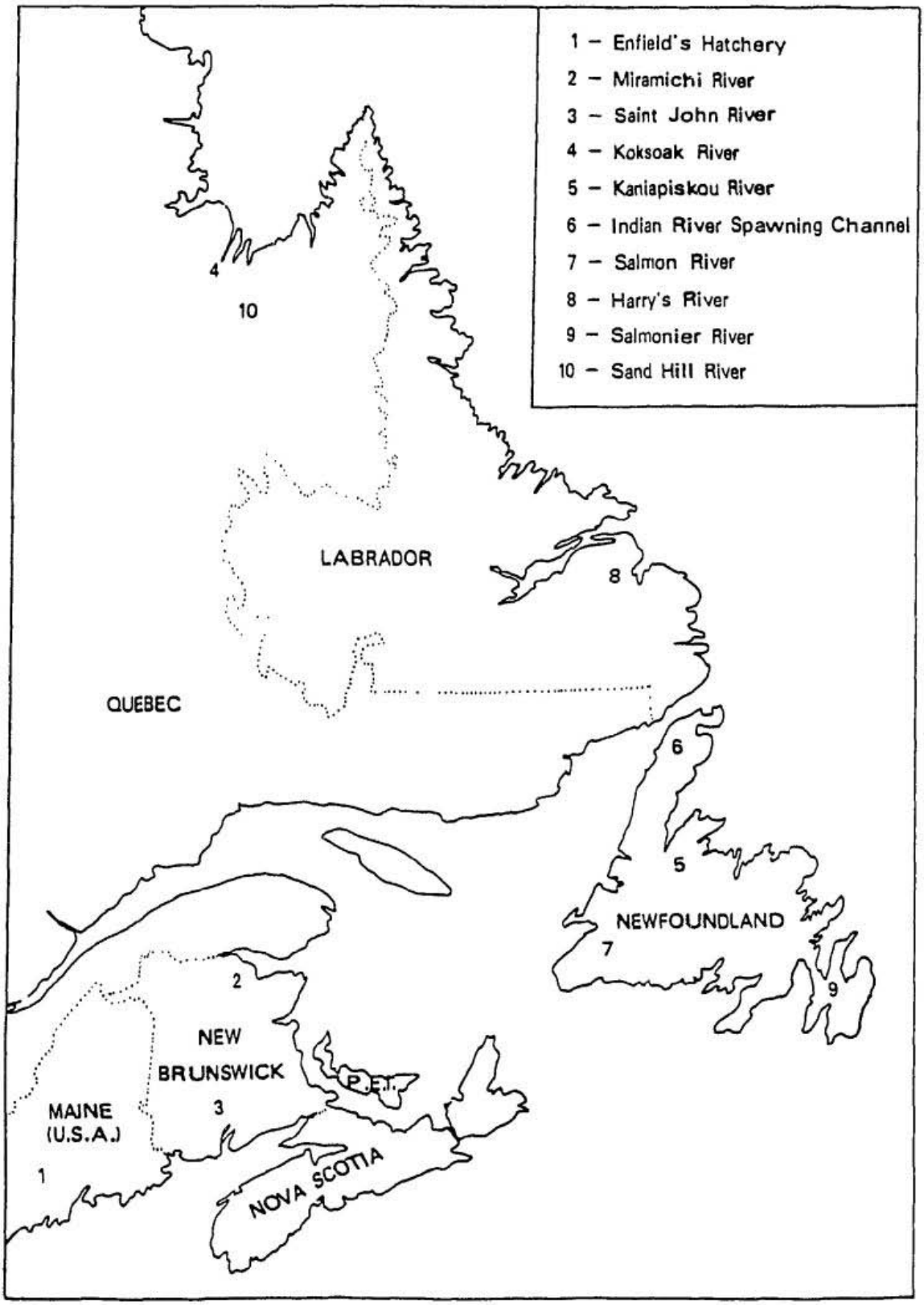


Figure 1.1b Locations of European Rivers

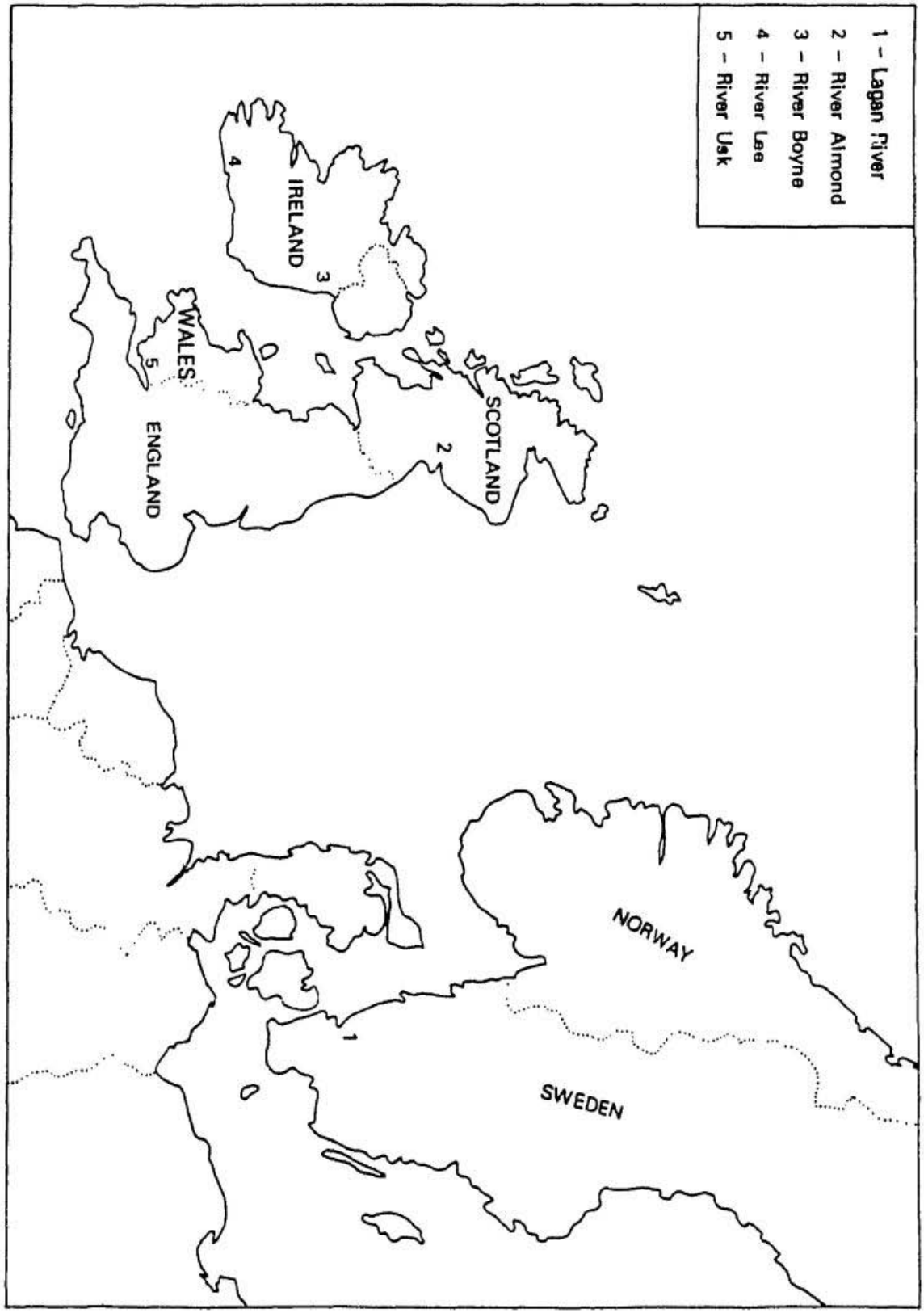
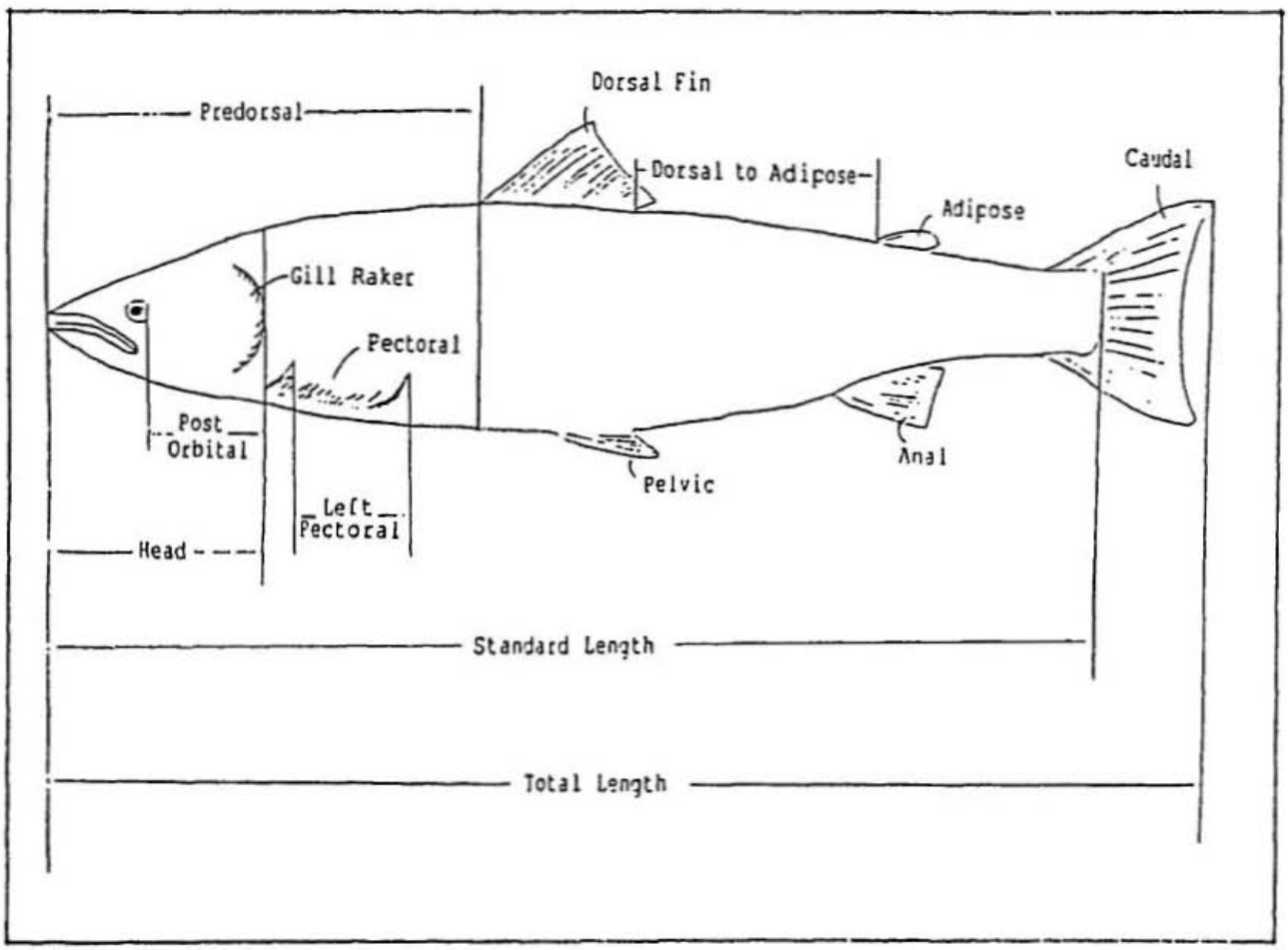


Figure 1.2 Measured Morphometric Variables



Chapter 2

EXPLORATORY ANALYSIS OF DATA

2.0 Introduction

Many statistical analyses assume that data consisting of more than one variable follow a multivariate normal distribution. One of the main reasons for this assumption is that the distributional results under normality are well known. However, there are situations where the normality assumptions may not be appropriate and in these cases transformation of the data is required prior to statistical analysis. If the underlying distribution is not normal and the analysis is done assuming normality, the results might be unreliable in certain cases. Thus, it is important to study the the distributional pattern of the data. With this in mind, the beginning of this chapter takes an initial look at the data. This is traditionally known as "Exploratory Data Analysis". Further, in Sections 2.1.1, 2.1.2 and 2.1.3, confirmatory analysis on the distributional pattern of the data is given.

The present analysis will be confined to the following three samples: specimens sampled from European rivers in 1969; specimens sampled from North American rivers in 1969; and those sampled from North American rivers in 1968. Only complete data will be used for this analysis, i.e., specimens for which all seven measurements are

available because missing observations virtually destroy morphometrics [Pimentel, p. 191 (1979)].

As a part of the exploratory data analysis, the data is examined for symmetry Box-plots are one of the appropriate graphical tools by which we may check symmetry. Boxplots can also help to identify the outliers in a data set. Specifically boxplots show the middle of a data set, from hinge to hinge, as a box with a "+" indicating the median (Hinges represent the upper and lower quartiles). The median can be defined as the middle observation in an ordered data series. The boxplot runs a solid line from each hinge to the corresponding extreme. At a glance, impressions can be made of the overall distribution, amount of spread, and symmetry of the data. Figures A2.1, A2.2 and A2.3 (Appendix A) show boxplots for all seven variables of European data sampled in 1969 and North American data sampled in 1969 and 1968 respectively. These boxplots are summarized in Sections 2.1, 2.2 and 2.3.

Some of these data series contain outliers, that is, values so high or low, that they stand out from the rest of the data. Values between the inner and outer fence are possible outliers, and are plotted with a "*". Values beyond the outer fence are probable outliers and are plotted with a "0". The inner and outer fence are defined as follows:

$$\begin{aligned} \text{inner fences} &= (\text{lower hinge}) - (1.5 \times (\text{H-spread})) \\ \text{and} &= (\text{upper hinge}) + (1.5 \times (\text{H-spread})) \end{aligned}$$

$$\begin{aligned} \text{outer fences} &= (\text{lower hinge}) - (3 \times (\text{H-spread})) \\ \text{and} &= (\text{upper hinge}) + (3 \times (\text{H-spread})) \end{aligned}$$

where H-spread = (upper hinge) - (lower hinge)

If a measurement is determined to be a probable or possible outlier, the whole observation (or record) is deleted from the data set. Note that in some cases valid data points may be dropped because they are atypical of the mass of data under analysis. However, because of the large sample size, this will not significantly affect the results of this particular analysis.

After the removal of outliers, there were 495 observations for European data sampled in 1969, 915 observations for North American data sampled in 1969 and 724 observations for North American data sampled in 1968. These sample sizes will be used for the remaining analysis.

Histograms were then displayed for each of seven variables for each of the three groups. The outliers were excluded while constructing the histogram and subsequently for the remainder of the analysis. The histograms are shown in Appendix A for all seven characters. The histograms for European data of 1969, North American data of 1969, and North American data if 1968 are displayed in figures A2.4, A2.5 and A2.6 respectively.

The histograms contained in figure A2.4 are summarized in Section 2.1.1. Similarly the histograms of figures A2.5 and A2.6 are summarized in Sections 2.1.2 and 2.1.3 respectively.

In the preceding analysis, graphical summaries of the data have been presented using relative frequency histograms and boxplots. Further analysis will investigate the data series using numerical summaries. For the seven variables in each group, the letter-value spreads H, E, D, C, B, A, Z, Y and X are recorded [Velleman and Hoaglin, (1981)]. The median,

M, splits an ordered data series in half. If the number of observations, n, is odd, the median, m, is found by the $\left(\frac{n+1}{2}\right)$th observation. If n is even, the median is the average of the $\left(\frac{n}{2}\right)$th and the $\left(\frac{n+2}{2}\right)$th observations.

The letter H denotes the hinges which are the summary values in the middle of each half of the data. They are about a quarter of the way in from each end of the ordered batch. Similarly, the letter E denotes the eighths and they are the middle values for the outer quarters of the data. These values are about an eighth of the way in from each end of the ordered batch. The pattern is continued for the letter-values D, C, B, A, Z, Y and X.

The difference between the lower hinge and upper hinge is known as the H-spread. Similarly, the E-spread is the difference between the lower eighth and the upper eighth, that is, the E-spread gives the range of the middle three-quarters of the data. The D-spread gives the range of the middle seven-eighths, and so on. These spreads are compared to the spreads for the normal, or Gaussian, distribution. The standard Gaussian spreads are: H-spread = 1.35, E-spread = 2.30, D-spread = 3.07, C-spread = 3.72, B-spread = 4.31, A-spread = 4.84, Z-spread = 5.32, Y-spread = 5.76 and X-spread = 6.18. The spreads of the data are compared with the Gaussian spreads by quotients of the spread values of the data to the Gaussian spread values. A trend in the quotients provides an indication of how the data depart from normality. If the quotients increase, the tails of the distribution are heavier than the tails of the Gaussian-shape. If the quotients shrink,

the tails of the data are lighter.

The average value of any two pair of letter values, called the mid-summary is also observed. Specifically, the average of the two hinges is called the mid-hinge; the average of the two eighths is called the mid-eighth, and so on. By observing a trend in the midsummaries, one can learn about the symmetry of the data. If the midsummaries become progressively larger, the data is skewed to the right. If they decrease steadily, the data is skewed to the left. Tables A2.1, A2.2 and A2.3 in Appendix A display the midsummaries, spreads and quotients for all seven characters for European data of 1969 and North American data of 1969 and 1968 respectively.

The variables from each group can be summarized by studying histograms (figures A2.4, A2.5 and A2.6), midsummaries, spreads and quotients (tables A2.1, A2.2 and A2.3). These summaries, both graphical and numerical, give indications about the distributional shape of the data.

2.1.1 Description of Variables for European Data Sampled in 1969

Total Length - This histogram (figure A2.4a, Appendix A) gives the impression of a bimodal distribution, that is, a distribution consisting of two peaks. The increasing values of the midsummaries indicates a slight skewness to the right of the data. The smaller second peak indicated in the histogram could be a reason for this shift. Also, the decreasing values of the quotients indicate a light-tailed distribution. There-fore, the normality of this

distribution is questionable.

Standard Length - The distributional shape of this variable is similar to the distribution of the variable total length. There are two peaks in the data and the increasing values of the midsummaries indicates a skewness to the right of the data. The quotients are also decreasing which indicates, as before, a light-tailed distribution. Again, the normality of this distribution is questionable.

Predorsal - The histogram (figure A2.4c, Appendix A) shows a concentration of the data toward the centre of the distribution. This indicates a light- tailed distribution which is verified by the decreasing quotient values. The mid-summary values show no indication of skewness. Thus, the distribution of this variable may be close in shape to the normal distribution.

Dorsal to Adipose - A bimodel distribution is observed similar to the distributional shape of variables total length and standard length. The slight increasing values of the mid-summaries indicates that the data are slightly skewed to the right. The decreasing quotient values indicate a light-tailed distribution. If there is any deviation from normality, it will be very small.

Head - The histogram for this data (figure A2.4e, Appendix A) also shows a concentration toward the centre of the distribution (similar to

the distribution of the variable predorsal). The decreasing quotient values indicate a light-tailed data series. The mid-summary values do not show any significant increasing or decreasing trend, therefore indicating a near symmetrical distribution with no skewness. As a result, this distribution can be considered as being close to normality.

Postorbital - The mid-summary values do not show any increasing or decreasing trend, indicating a symmetrical distribution. The quotient values show a slight decreasing trend for the H, E and D spreads, but remain relatively constant for the remainder of the spread values. This may indicate a slight light-tailed distribution. Therefore, this distribution can also be considered being close to normal.

Left Pectoral - The information obtained from the histogram (figure A2.4g, Appendix A) indicates that this data batch approximates normality better than any of the previous variables. The slightly increasing mid-summary values suggest that there is a small skewness to the right. The quotient values remain relatively constant, indicating normal tails. Therefore, this data series approximates the normal distribution quite well.

In summary, it is seen that four variables out of seven approximately follow the normal distribution. Most questionable are the variables total length, standard length and dorsal to adipose. These characters have bimodal distributions, are skewed to slightly to the right and

may be light-tailed in their distributional shape. However, these deviations from normality are not extreme. Further review (Sections 2.2 and 2.3) will show that these deviations will not significantly influence the analysis.

2.1.2 Description of Variables for North American Data Sampled in 1969

Total Length - The histogram for the data series (figure A2.5a, Appendix A) give no indication of skewness, but the mid-summary values show a decreasing trend, indicating that the data is skewed to the left. The quotient values are constant except for the A, Z, Y and X letter values, which show an increasing trend, indicating the possibility of a heavy-tailed distribution. However, no strong deviations from normality are apparant.

Standard Length - This distribution behaves similar to the distribution for total length. The mid-summary values show a decreasing trend, indicating skewness to the left, but there is no evidence of this from the histogram. The quotient values are also constant except for the A, Z, Y or X letter values. Therefore, the distributional shape of this variable is close to normality.

Predorsal - Again, this distribution has similar qualities to the distribution of the previous two variables. The mid-summary values are decreasing, indicating a skewness to the left, and the quotient values

remain constant except for the A, Z, Y and X spreads. This variable has a distributional shape which is close to normality.

Dorsal to Adipose - The mid-summary values for this distribution are relatively constant, decreasing a little for the last few letter values. However, the histogram (figure A2.5d, Appendix A) does not indicate any skewness and the quotient values show an increasing trend, maybe indicating a heavy-tailed distribution. Therefore, this indicates that the distribution follows normality relatively well.

Head - The distributional shape of data is again similar to previous variables in this group. Decreasing mid-summary values may indicate a slight skewness to the left. Increasing quotient values may indicate a heavy-tailed distribution. However, these deviations are very slight, indicating that the distribution is close to normal.

Postorbital - Once again, this distribution has similar properties. Decreasing mid-summary values indicate a slight skewness to the left. This slight skewness can be detected in the histogram. The increasing quotient values also indicate a heavy-tailed distribution. However, despite these slight deviations, it can be said that the distribution is relatively close to normal.

Left Pectoral - This is another distribution with similar characteristics. Skewness to the left is indicated by the decreasing mid-summary values although the histogram looks to be symmetrical. A trend does not exist for the quotient values except for the A, Z, Y and X letter values. Therefore, the distribution is approximately normal.

In summary, it can be seen that all variables for this particular group have similar distributional properties. All variables show possible signs of a skewness to the left, however, if a skewness exists, it is very slight. Another feature common amongst these variables is a heavy-tailed distribution. Again, this is not an extreme deviation. Therefore, all variables in this particular group can be said to approximate a normal distribution.

2.1.3 Description of Variables for North American Data Sampled in 1968

Total Length - The distributional shape of the histogram (figure A2.6a, Appendix A) does not show any deviations from normality. The quotient values do not show an increasing or decreasing trend but the mid-summary values do show an increasing trend, indicating a possible skewness to the right. However, it is very minimal since it cannot be detected from the histogram. Therefore, this distribution seems close to normal.

Standard Length - The mid-summary values as well as the quotient values show no significant trend. No skewness is indicated by the histogram.

Therefore, it is safe to assume that this distribution is normal.

Predorsal - Again, no skewness is present in the histogram. The mid-summary values are constant but the quotient values are decreasing very slightly, which may indicate a light-tailed distribution. However, these deviations are very small which leads one to believe that the distributional shape is normal.

Dorsal to Adipose - The quotient values do not show a trend for this distribution. However, the mid-summary values show a slight increasing trend, indicating a distribution that is skewed to the right, and the histogram does not show any skewness at all. Therefore, any skewness present in this distribution is very minimal. Thus, indications are that this distribution is normal.

Head - Although the histogram shows what appears to be a skewed distribution, there isn't any indication of this from the mid-summaries. Also, the quotient values do not show an increasing or decreasing trend. Thus, this distribution can be assumed to be approximately normal.

Postorbital - Once again, there is no trend in the mid-summary values and quotient values and there is no indication of skewness in the histogram. Therefore, this distribution is close to normal.

Left Pectoral - Again, there is not an indication of skewness from the histogram and mid-summaries and the quotient values show neither an increasing or decreaasing trend. Thus this distribution follows an approximate normal shape.

In summary, it is seen that all variables of this group follow the normal distribution. There are no indications of a light-tailed or heavy-tailed data series. Any variables which were shown to have a skewed distribution, were skewed very slightly.

So far, the distributions for each of the seven variables sampled from European rivers in 1969 and North American rivers in 1968 and 1969 have been studied. Most of the variables were found to satisfy the property of the normal distribution. The possible exceptions are total length, standard length and dorsal to adipose variables sampled from European rivers.

However, marginal normality does not necessarily imply the joint multivariate normality of all characters (Anderson, 1958), although, it gives a good indication. In the following section, the joint distributional features of each group is studied.

Note that although the variable total length was included in this section, it will not be included in the following analysis. Recall from Section 1.3 the definitions of the measurements total length and standard length. Total length is the greatest dimension between the tip of the specimen's snout and the furthest tip of the caudal fin

measured in a straight line. Standard length is the distance between the tip of the snout back to the end of the vertebral column. Since these two variables are very similar measurements (their distributional patterns are also similar), one of the two variables can be dropped. On some specimens, the caudal fin may be ragged or torn, thus giving an inaccurate measurement for the variable total length. Therefore, standard length was selected over total length.

2.2 Numerical Test for Normality - the Univariate Case

One of the assumptions in attempting a discriminant analysis is that the variables in a group follow a multivariate normal distribution. If the data do not follow a multivariate distribution, then transformation of data is performed to obtain a normal data set.

Box and Cox (1964) proposed a method of shifted-power transformation of a single non-negative variate X to Y where

$$Y = \begin{cases} (X^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \ell n\ X & \lambda = 0 \end{cases}$$

More extensive computations would be involved in considering analogues of the more general class of shifted power transformation, that is, X may be replaced with $X + \varepsilon$ in the above. Assuming that (ε,λ) is the pair yielding normality, the MLE of ε and λ is obtained. Then

$$\ell n[L_{max}(\hat{\varepsilon},\hat{\lambda})] - \ell n[L_{max}(\varepsilon,\lambda)] \leq \frac{1}{2}\chi^2_{2,\alpha}$$

where $L_{max}(\varepsilon,\lambda)$ is the maximum likelihood estimate (MLE) of ε and λ, and $\chi^2_{2,\alpha}$ is the upper α-point of χ^2 with 2 degrees of freedom. If this region contains $\lambda = 1$, the hypothesis of normality is accepted.

This idea was used to determine which characters, if any, deviated from normality. Only European data needed to be tested and initially each character was tested for univariate normality (the testing of multivariate normality is dealt with in the next section). The $\ell n[Lmax(\lambda)]$ was calculated for $\lambda = 0.00, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 2.50$ and 3.00 using equation 2.3.1 of Section 2.3. The maximum value of $\ell n[Lmax(\lambda)]$ determined λ_i, the coefficient for transformation to normality. The following table show the results:

Table 2.1: Values of $\ell n[Lmax(\hat{\lambda})]$

λ	Standard Length	Predorsal	Dorsal	Head	Postorbital	Left Pectoral
0.00	-1408.7278	-979.1898	-814.9882	-574.1053	-306.9319	-370.4496*
0.25	-1407.7521	-978.1089	-809.3297	-571.4026*	-300.1801	-371.1551
0.50	-1407.1332	-977.0673	-806.4337	-571.5027	-298.9677	-371.6977
0.75	-1406.7924*	-976.6109*	-804.1299	-572.0068	-298.8820*	-372.8702
1.00	-1407.1436	-976.8740	-803.0921*	-572.4505	-299.1438	-374.0129
1.50	-1410.0227	-979.1631	-804.3463	-574.5005	-301.0980	-377.5385
2.00	-1415.7861	-983.7908	-810.0248	-577.9480	-304.9042	-382.3832
2.50	-1424.4146	-990.7092	-820.0454	-582.7810	-310.5352	-388.5293
3.00	-1435.8789	-999.8643	-834.3308	-588.9846	-317.9702	-395.9617
X^2	0.7024	0.5262	0.0000	2.0958	0.5236	7.1266**

* maximum value of $\ell n[(Lmax(\hat{\lambda})]$

** significant at $\alpha = 0.01$

As seen from table 2.1, the only transformation significant in testing for normality was $\lambda = 0$ for the variable left pectoral (this is the natural log transformation).

2.3 Numerical Test for Normality - the Multivariate Case

Andrews, Gnanadesikan and Warner (1971) extended the univariate transformation on the responses to the multivariate case. Let $\underset{\sim}{X}$ be a $p \times N$ vector where p is the number of variables and N is the number of observations, and each element $X_{ij} > 0$. Let $\underset{\sim}{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ be a vector of powers defined by

$$Y_{ij} = \begin{cases} (X_{ij}^{(\lambda_i)} - 1)/\lambda_i & \lambda_i \neq 0 \\ \ell n(X_{ij}) & \lambda_i = 0 \end{cases}.$$

Then the transformed data matrix may be described as a p-variate normal model with a mean vector $\underset{\sim}{\mu}$ and a covariance matrix $\underset{\sim}{\Sigma}$. Consequently, it can be shown that

$$\ell n[\text{Lmax}(\underset{\sim}{\lambda})] = -\frac{N}{2} \ell n \left|\underset{\sim}{\hat{\Sigma}}\right| + \left[\sum_{i=1}^{p} (\lambda_i - 1) \sum_{j=1}^{N} \ell n(X_{ij})\right] \quad (2.3.1)$$

One can find $\underset{\sim}{\hat{\lambda}}$ by maximizing $\log[\text{Lmax}(\underset{\sim}{\lambda})]$. The hypothesis of normality, i.e. $\underset{\sim}{\lambda} = 1$, may be tested based on the statistic

$$2\{\ell n[\text{Lmax}(\underset{\sim}{\hat{\lambda}}) - \ell n[\text{Lmax}(j_p)]\}, \; j_p = (1, 1, \ldots, 1)' \quad (2.3.2)$$

which is asymptotically distributed as χ^2 with p degrees of freedom. As seen in the previous section, the only questionable variable in testing for univariate normality was the variable left pectoral. All variables

were tested for multivariate normality where λ_i, i = 1, ..., 5, the coefficients for transformation to normality corresponding to the variables standard length, predorsal, dorsal, head and postorbital remain constant. The coefficient corresponding to the variable left pectoral, λ_6, varied from 0.00 up to 3.00 as before. The following table summarizes the results:

Table 2.2: Values of $\ell n[Lmax(\hat{\underset{\sim}{\lambda}})]$

λ	$\ell n[Lmax(\hat{\underset{\sim}{\lambda}})]$
0.00	-1739.0598
0.25	-1731.1512*
0.50	-1733.7260
0.75	-1733.8470
1.00	-1734.2043
1.50	-1737.7616
2.00	-1738.9448
2.50	-1753.7422
3.00	-1761.8872

*maximum value of $\ell n[Lmax(\hat{\underset{\sim}{\lambda}})]$

Here, $X^2 = 6.1062$ and $\chi^2_{.1,5} = 9.236$. Therefore, the test is not significant at the 10% level of significance and it is concluded that there is no significant departure from multivariate normality. No transformations will be necessary in the remainder of the analysis.

Chapter 3

DISCRIMINANT ANALYSIS

3.0 Introduction

As mentioned in Chapter 1, it is very important to estimate the proportions of North American and European Atlantic salmon in the population of salmon the fishery at West Greenland. To estimate these proportions one requires the identification of specimens of unknown origin. By identification of a specimen, it is specifically meant that a salmon whose home river is in Europe should be identified as a European origin salmon, and a salmon whose home river is in North America should be identified as a North American origin salmon. Once the identity of the specimen is determined, the proportions of North American to European salmon off West Greenland can be estimated.

The importance of the above identifications to estimate proportions of North America and European salmon is well discussed in the literature. For example, Ritter, Marshall, Reddin and Doubleday (1980) assessed the impact of the West Greenland fishery on stocks and catches in North America. At that time, for each tonne of North American origin salmon caught at West Greenland, the loss to homewater stocks was estimated to range from 1.70 to 2.42 tonnes. Similarly, the loss to homewater catches was projected to range from 1.58 to 2.11 tonnes.

As a result, the yield increase to all fisheries with any reduction in catch of North American origin salmon at Greenland was estimated to range from 58% to 111%. Their assessment indicated that the exploitation of salmon at West Greenland was resulting in a reduced yield to all fisheries in homewaters.

Since salmon is economically an important species to many countries, the smolt data sampled from European rivers in 1969 and data sampled from North American rivers in 1968 and 1969 are chosen. Thus, any salmon caught at West Greenland can be sampled to study their identification through the classification technique.

Note that there is vast literature on classification techniques. In order to classify an observation into one of the populations, in an early paper, Fisher (1936) suggested, as a basis for classification decisions, the use of a discriminant function linear in the components of the observations. Other bases for classification have included likelihood ratio tests (Anderson, 1958), information theory (Kullback, 1959), and Bayesian techniques (Geisser, 1964). In all cases, sampling theories have been considered under the assumption that the populations involved are multivariate normal. As the six variables: standard length, predorsal, dorsal, head, postorbital and left pectoral were found to follow the multivariate distribution, for the classification problem, classical methods of discrimination based on the multivariate normal distribution can be utilized.

3.1 Eliminating the Effect of Size

Morphometric variables, that is, variables that describe body form, are measures of the absolute sizes of body parts. Reist (1985) reported that for specimens in which determinate growth exists, there is a variation in absolute size within and between groups of specimens. Furthermore, any heterogeneity in size across samples will result in heterogeneity in shape. Thus, the differences in shape may be the result of size variation and may not reflect any new information. Alternatively, the shape of the specimen at a particular size may vary across samples and thus reflect a difference between specimens. Therefore, comparison of samples should be in terms of variables free from the effect of size.

Different methods have been proposed to eliminate the effect of size in comparing samples. One technique widely used is the creation of a ratio between each of the p variables, $(X_1, X_2, \ldots, X_p)$, and some standard measure, Z (standard length in this analysis). The shape estimate for the jth specimen of the ith variable in a single population would be:

$$Y_{ij} = \frac{X_{ij}}{Z_{ij}}$$

However, this ratio method has come under criticism for its undesirable statistical properties, for example, we refer to Atchley, Gaskins and Anderson (1976), but its use still continues (cf. Mosimann and James, 1979; Shaklee and Tamaru, 1981; Wilk et. al., 1980). Furthermore, ratios do not completely remove the influence of size variation

from the data (Albrecht, 1978; Atchley, Gaskins and Anderson, 1976; Dodson, 1978).

Another technique used in the adjustment of size variation in the data is the regression technique. The appropriate regression equation is:

$$Y = X - \beta(Z - \bar{Z}) \qquad [3.1.1]$$

each of the p variables, X is the original unadjusted measurement, Z is the standard measure of the invidiual, $\bar{Z}$ is the grand mean of the standard length across all individuals, and β is the slope of the relationship between X and Z. This technique enables one to predict a specimen's size for a particular variable given that the specimen has a mean standard length. This technique can be used to remove the effect of the standard length for the remaining five variables of the analysis. These five "adjusted" variables will be used in the discrimination analysis.

Let X_1, X_2, X_3, X_4 and X_5 represent the variables predorsal length, dorsal to adipose, head length, postorbital length and left pectoral length respectively and let Z represent the covariate standard length. By 3.1.1

$$\underset{\sim}{Y} = \underset{\sim}{X} - \underset{\sim}{B}(Z - \bar{Z}), \text{ where } \underset{\sim}{Y} = [Y_1, Y_2, Y_3, Y_4, Y_5]' ,$$
$$\underset{\sim}{X} = [X_1, X_2, X_3, X_4, X_5]' , \qquad [3.1.2]$$
$$\text{and } \underset{\sim}{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]' .$$

The covariance matrix of X_1, X_2, X_3, X_4, X_5 and Z is partitioned as follows:

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma'_{xz} & \Sigma_{zz} \end{bmatrix}$$

where

$$\Sigma_{xx} = \begin{bmatrix} \mathrm{Var}(X_1) & & & & \\ \mathrm{Cov}(X_1,X_2) & \mathrm{Var}(X_2) & & & \\ \mathrm{Cov}(X_1,X_3) & \mathrm{Cov}(X_2,X_3) & \mathrm{Var}(X_3) & & \\ \mathrm{Cov}(X_1,X_4) & \mathrm{Cov}(X_2,X_4) & \mathrm{Cov}(X_3,X_4) & \mathrm{Var}(X_4) & \\ \mathrm{Cov}(X_1,X_5) & \mathrm{Cov}(X_2,X_5) & \mathrm{Cov}(X_3,X_5) & \mathrm{Cov}(X_4,X_5) & \mathrm{Var}(X_5) \end{bmatrix},$$

$$\Sigma_{xz} = \begin{bmatrix} \mathrm{Cov}(X_1,Z) \\ \mathrm{Cov}(X_2,Z) \\ \mathrm{Cov}(X_3,Z) \\ \mathrm{Cov}(X_4,Z) \\ \mathrm{Cov}(X_5,Z) \end{bmatrix} \qquad \text{and} \qquad \Sigma_{zz} = \mathrm{Var}(Z) \ .$$

Using the above notations, 3.1.1 can be rewritten as

$$\underset{\sim}{Y} = \underset{\sim}{X} - \beta_i (Z - \bar{Z}) \ .$$

Since $\beta_i = \dfrac{\mathrm{Cov}(X_i,Z)}{\mathrm{Var}(Z)}$ then, from 3.1.2

$$\underset{\sim}{Y} = \underset{\sim}{X} - \Sigma_{xz}\Sigma_{zz}^{-1}(Z - \bar{Z}) \ . \qquad [3.1.3]$$

As it has been demonstrated in the last chapter that $[\underset{\sim}{X}, Z]$ has a six dimensional multivariate normal distribution, one writes

$$\begin{pmatrix} \underset{\sim}{X} \\ Z \end{pmatrix} \sim N_6\left[\begin{pmatrix} \underset{\sim}{\mu_x} \\ \mu_z \end{pmatrix} , \ \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}\right] \ . \qquad [3.1.4]$$

Consequently,

$$X|Z=z \sim N_5(\mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(Z - \mu_z), \ \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}) \ ,$$

i.e. $$E(X|Z=z) = \mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(Z - \mu_z) \ , \qquad [3.1.5]$$

and $$V(X|Z=z) = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx} \ . \qquad [3.1.6]$$

By using 3.1.5 and 3.1.6 in 3.1.3,

$$E(Y|Z) = \mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(\bar{Z} - \mu_z)$$

and $$V(Y|Z) = V(X|Z) \ .$$

These conditional means and conditional variances will be estimated by $\bar{X}$ and $S_{xx} - S_{xz}S_{zz}^{-1}S_{zx}$ respectively where $S_{xx} = \sum_i (x_i - \bar{x})(x_i - \bar{x})'$ and $S_{xz} = \sum_i (x_i - \bar{x})(z_i - \bar{z})'$.

In summary, the analysis of covariance will adjust each of the variables for each group to the overall mean standard length according to the formula:

$$\underset{\sim}{Y} = \underset{\sim}{X} - \underset{\sim}{\beta}(Z - \bar{Z})$$

where $\underset{\sim}{X}$ is the original vector, Z is the standard length of each individual specimen and $\underset{\sim}{Y}$ is the covariate of the adjusted variables.

In the following, the conditional variables $\underset{\sim}{Y}$ where Z is given, are considered, such that $Z = z$.

3.2 Testing the Differences Between Groups

Prior to discriminant analysis, it is necessary to test whether or not a significant separation exists between any two groups. That is, $H_0: \underset{\sim}{\mu}_{1Y} - \underset{\sim}{\mu}_{2Y} = 0$ is tested against $H_1: \underset{\sim}{\mu}_{1Y} - \underset{\sim}{\mu}_{2Y} \neq 0$, where $\underset{\sim}{\mu}_{1Y}$ is the mean of the five conditional Y variables of the first population and $\underset{\sim}{\mu}_{2Y}$ is the mean of the five conditional Y variables of the second population. Let $\underset{\sim}{Y}_{ij}$ be the five dimensional variables for the jth observation in the ith population, and let $\bar{Y}_i$ be the sample mean vector for the ith sample. Then the above hypothesis may be tested by the Mahalanobis generalized sample squared distance, D^2:

$$D^2 = (\bar{\underset{\sim}{Y}}_1 - \bar{\underset{\sim}{Y}}_2)' S_p^{-1} (\bar{\underset{\sim}{Y}}_1 - \bar{\underset{\sim}{Y}}_2) \qquad [3.2.1]$$

where

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \qquad [3.2.2]$$

and where

$$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)(Y_{1j} - \bar{Y}_1)' , \qquad [3.2.3]$$

$$S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)(Y_{2j} - \bar{Y}_2)' , \qquad [3.2.4]$$

and n_1 and n_2 are the respective sample sizes. The value of S_p is often referred to as the pooled variance-covariance matrix. One can now use the distribution of D^2 to test if there are significant differences between the two groups. The statistic (often referred to as Hotelling's T^2 statistic) is given by

$$W = \left[\frac{(n_1 + n_2 - k - 1)}{(n_1 + n_2 - 2)k}\right]\left[\frac{n_1 n_2}{n_1 + n_2}\right] D^2 \qquad [3.2.5]$$

where D^2 is as in [3.2.1] and k is the number of variables. It is well known that (cf. Johnson and Wichern, 1982) under H_0, $W \sim F(k, n_1 + n_2 - k - 1)$. The larger the value of D^2, the greater the distance between the groups, and as a result, the large value of W would lead to the rejection of the null hypothesis.

W and the corresponding p-values are calculated in testing the significance of the separation of any two populations (pairwise) considered in the study. The results in tabular form are shown in Appendix A. A brief description of these results is also given in the following.

In Table A3.1, Appendix A, the separation between the two populations, North America and Europe, is examined. The value of W is very large, yielding a very small p-value. Thus, the populations are well separated. Similar comparisons have been made between the five rivers of Europe in Table A3.2. All values of W were large implying that a selected river is well-seperated statistically from any other river. Similarly Tables A3.3 and A3.4 show the same results for the six rivers of North America (1968) and the eight rivers of North America (1969) respectively.

In summary, it has been determined in this section that all pairwise populations are significantly different about the five conditional variables. Group 1 will be referred to as the whole data set consisting of North American (1969) and European salmon. Similarly, group 2 will be referred to as the European data set and groups 3 and 4 as the North American (1969) and North American (1968) data sets respectively. Since any two rivers under any of the four groups are well seperated, a discriminant analysis can be performed in order to assign a specimen to its population.

Note, however, that although the rivers under a group are well-seperated, they still may overlap each other to a certain extent. Consequently, there may be errors in assigning the specimens. This is a misclassification problem which will be discussed in Section 3.6.

3.3 Determination of Discriminant Functions

The general underlying theory for the determination of discriminating functions will be as follows: Let $y_0 = (y_{10}, y_{20}, y_{30}, y_{40}, y_{50})$ be an observation which may arise due to one of the populations: $\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_g$, where $g = 2$ for group 1, $g = 5$ for group 2, $g = 8$ for group 3 and $g = 6$ for group 4.

As shown in Chapter 2, the samples can be considered to be multivariate normal. Therefore, it can be considered without any loss of information that $\pi_i \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}_i)$, where μ_i is the population vector mean and Σ_i is the population variance-covariance matrix. The appropriate classification criterion, under the assumption of equal misclassifi-

cation costs, for assigning $\underset{\sim}{y}_0$ to one of the π_i's is given by:

Allocate $\underset{\sim}{y}_0$ to π_k if

$$\ell n\ p_k f_k(Y) = \ell n(p_k) - \left(\frac{p}{2}\right)\ell n(2\pi) - \frac{1}{2}\ell n|\underset{\sim}{\Sigma}_k| - \frac{1}{2}(\underset{\sim}{Y}_0 - \underset{\sim}{\mu}_k)'\underset{\sim}{\Sigma}_k^{-1}(\underset{\sim}{Y}_0 - \underset{\sim}{\mu}_k)$$

$$= \text{maximum value of } p_i f_i(Y) \qquad [3.3.1]$$

$$\text{for } i = 1, 2, \ldots, g ,$$

and where p_i = prior probability of the observation being contained in the ith population and $f_i(Y)$ are multivariate normal densities (ref. Johnson and Wichern, 1982).

The constant $\left(\frac{p}{2}\right)\ell n(2\pi)$ can be ignored in equation 3.3.1 since it is equal for all populations. The quadratic discrimination score for the ith population is now defined as:

$$d_i^Q(\underset{\sim}{Y}) = -\frac{1}{2}\ell n|\underset{\sim}{\Sigma}_i| - \frac{1}{2}(\underset{\sim}{Y}_0 - \underset{\sim}{\mu}_i)'\underset{\sim}{\Sigma}_i^{-1}(\underset{\sim}{Y}_0 - \underset{\sim}{\mu}_i) + \ell n(p)_i \qquad [3.3.2]$$

$$\text{for } i = 1, 2, \ldots, g .$$

The quadratic score, $d_i^Q(\underset{\sim}{Y})$, is composed of contributions from the generalized variance $|\underset{\sim}{\Sigma}_i|$, the prior probability p_i, and the squared distance from $\underset{\sim}{Y}$ to the population mean $\underset{\sim}{\mu}_i$. Using discriminant scores, the classification rule of [3.3.2] becomes the following:

Allocate $\underset{\sim}{Y}_0$ to π_k if $d_k^Q(\underset{\sim}{Y}) = \text{Max}[d_1^Q(\underset{\sim}{Y}), d_2^Q(\underset{\sim}{Y}), \ldots, d_g^Q(\underset{\sim}{Y})]$ where $d_i^Q(\underset{\sim}{Y})$ is given by equation [3.3.2].

This can be referred to as the Minimum Total Probability of Misclassification Rule for Normal Populations.

Note that $\underset{\sim}{\mu}_i$ and $\underset{\sim}{\Sigma}_i$ in [3.3.2] are unknown. In order to compute all necessary discrimination scores, the following estimates are used:

$$\hat{\underset{\sim}{\mu}}_i = \bar{\underset{\sim}{Y}}_i = \bar{\underset{\sim}{X}}_i - \underset{\sim}{S}_{xz}^{(i)} \underset{\sim}{S}_{zz}^{(i)^{-1}} (\underset{\sim}{Z}_i - \bar{\underset{\sim}{Z}}_i)$$

$$\hat{\underset{\sim}{\Sigma}}_i = \underset{\sim}{S}_i = \underset{\sim}{S}_y^{(i)} = \underset{\sim}{S}_{xx}^{(i)} - \underset{\sim}{S}_{xz}^{(i)} \underset{\sim}{S}_{zz}^{(i)^{-1}} \underset{\sim}{S}_{zx}^{(i)}$$

where $\bar{\underset{\sim}{Y}}_i$ and $\underset{\sim}{S}_i$ are the 5 X 1 sample mean vector and 5 X 5 sample covariance matrix respectively. The estimate of the quadratic discrimination score $\hat{d}_i^Q(\underset{\sim}{Y})$ is then:

$$\hat{d}_i^Q(\underset{\sim}{Y}) = -\frac{1}{2}\ell n|\underset{\sim}{S}_i| - \frac{1}{2}(\underset{\sim}{Y}_0 - \bar{\underset{\sim}{Y}}_i)' \underset{\sim}{S}_i^{-1} (\underset{\sim}{Y}_0 - \bar{\underset{\sim}{Y}}_i) + \ell n(p_i) \quad [3.3.3]$$

and the classification rule based on the sample is as follows:

Allocate $\underset{\sim}{Y}_0$ to π_k if $\hat{d}_k^Q(\underset{\sim}{Y}) = \text{Max}[\hat{d}_1^Q(\underset{\sim}{Y}), \hat{d}_2^Q(\underset{\sim}{Y}), \ldots, \hat{d}_g^Q(\underset{\sim}{Y})]$ where $\hat{d}_i^Q(\underset{\sim}{Y})$ is given by equation 3.3.3.

In summary, given a vector of observations of a specimen coming from an unknown population, and given g number of populations to choose from, $\underset{\sim}{Y}_0$ can be substituted into each of the g equations. If the kth equation gives the largest result, the specimen belonging to these particular observations should originate from the kth population. However, there is always the chance of misclassification, that is, concluding that a specimen belongs to a certain population when, in reality, it belongs to some other population. Also, for developing proportions, ie. the West Greenland fishery, there may be a problem with the error rate. Because fish are unclassed and because the number from each group are not necessarily equal, the proportions

developed from a given discriminant analysis may be biased. This misclassification problem will be discussed in Section 3.4.

3.3.1 Selection of Discriminating Variables: Considerations

In doing different studies and analyses, one may encounter several potential discriminating variables but may be uncertain whether all of them are valuable and necessary. In these situations, one or more of the variables may be poor discriminators because one or more of the means may be relatively "close". Also, two or more of the discriminating variables may be individually good discriminators, but may share the same discriminating information. Even though they may be good discriminators in a multivariate analysis, they do not contribute to a multivariate analysis because their unique characteristics are insufficient.

One way to eliminate unnecessary variables is by using a stepwise procedure to select the most important variables. There are three ways in which this can be done. The first method is a forward stepwise procedure. This procedure begins by selecting the individual variable which provides the best univariate discrimination. (This can be determined on the basis of several well-known criteria which will be covered in the next section.) The procedure then pairs this first variable with each of the remaining variables, one at a time, until a combination is found which produces the best discrimination. The procedure then goes on to combine this pair with each of the remaining variables until a combination of three is found which produces the greatest discrimination.

This procedure continues until all possible variables have been selected or the remaining variables do not contribute enough to the discriminating power.

The second method is a backward stepwise procedure. This procedure works in a backward direction in which all variables are initially included, and then the worst variable is cast out at each step.

Thirdly, these two procedures can be combined. This involves a forward selection procedure with each step starting with a review of the variables previously selected. If any of these variables no longer makes a sufficient contribution to the discrimination, then that variable is cast out, although it will be eligible to be selected again at any future step.

As the last procedure clearly has the advantage over the other two, it will be used in the selection of discriminating variables for this analysis.

Note that in order to choose the best solution of discriminating variables, one would have to test all possible combinations (all possible pairs, all possible combinations of three, etc.). Such testing would be very costly and time consuming. Thus, such testing is not attempted here.

Stepwise procedures used in a discriminant analysis must enter and remove variables one at a time, selecting them on the basis of certain criteria. There are several well-known criterion, for example, Wilk's lambda, Rao's V, Mahalanobis squared distance between closest groups,

Between-groups-F, and Minimizing Residual Variance.

The criteria chosen in this analysis will be Wilk's lambda. The reason is that this criterion takes into account both the differences between groups and the homogeneity within groups. Unlike other selection criteria, a variable which increases homogeneity without changing the separation between group centroids may be selected over a variable which increases separation without changing homogeneity. Here, Wilk's lambda, denoted by Λ, is given by:

$$\Lambda = \frac{|W|}{|T|}$$

where

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_{i\cdot})(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_{i\cdot})'$$

$$T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_{\cdot\cdot})(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_{\cdot\cdot})'$$

3.3.2 Discriminating Variables for the Salmon Data

To see how the selection technique described in the previous section works, a detailed explanation will be given for the stepwise discriminant analysis involving the two groups of salmon sampled from European and North American rivers in 1969. Results of other stepwise discriminant analysis will be given without any discussion.

The five variables taken into consideration are: Predorsal (PREDOR), Dorsal to Adipose (DORS), Head (HEAD), Postorbital (POSTOR),

and Left Pectoral (LFTPECT), all free from the effect of Standard Length. Before a variable is to be tested on the selection criterion, it must pass certain minimum conditions. These conditions are a tolerance test to assure computational accuracy; a partial F statistic to assure that the increased discrimination exceeds Λ ; and a check of the list of variables already entered to determine if any should be deleted.

Tolerance: This test is designed to preserve computational accuracy. The tolerance of a variable not yet selected is one minus the squared multiple correlation between that variable and all other variables already entered. The correlations are based on the within-group correlation matrix.

F-to-Enter: This is a partial multivariate F statistic which takes into account the discrimination achieved by the other variables already entered and tests the additional discrimination introduced by the variable being considered. If the F is small, it is not desirable to enter this variable because it will not add enough to the overall discrimination

F-to-Remove: This is also a partial multivariate F-statistic, but it tests the significance of the decrease in discrimination should that variable be removed from the variables already selected. This test is done at the beginning of each step to see if there are any variables which no longer make a sufficiently large contribution to discrimination. A variable that was a good choice earlier may not be valuable now because other variables could have been entered that

duplicate its contribution.

The results of the stepwise procedure are recorded in Table 3.1. On the first step, the tolerance level is always 1.0 because no variables have been entered and the F-to-enter corresponds to the univariate F-statistic. The fifth column gives the values for Wilk's lambda among which the smallest is selected. The value 0.27907 is produced by the variable PREDOR and the p-value of F-to-remove is 0.0000, which is less than 0.01. This is the first entry at step 1. Notice here that the variable PREDOR has an F-to-remove significance of 0.0000. (Recall that the F-to-remove is a partial F for the discrimination added by PREDOR after all other variables has created as much discrimination as possible. In this case there are no other variables.) Since this p-value is less than 0.01, it stays in and another variable is selected from the four remaining variables. At this stage, all relevant statistics are usually computed, taking into account that PREDOR has already been entered. Now the tolerance is less than one since it represents one minus the squared correlation between PREDOR and the respective variable. The F-to-enter is now the partial F for the discrimination added by the respective variable after PREDOR has created as much discrimination as possible. Thus the smallest Wilk's lambda is 0.16388 produced by the variable DORS, and since the p-value of the F-to-enter is 0.0000, the variable DORS is entered at this step.

In step 2, PREDOR and DORS are tested for removal, and both stay in since the p-value of the F-to-remove for both variables is 0.0000. The variable HEAD is now entered since it has the smallest

Wilk's lambda (0.12289) and the p-value for the F-to-enter is 0.0000. The remaining steps proceed in a similar fashion until all the variables have been entered that meet the requirements. Note that for this analysis, all variables were entered, so all variables will be used in determining the classification functions.

Table 3.1

Entry Statistics for Stepwise Selection[@]

(North America (1969) vs Europe (1969))

	Variable	Tolerance	Significance F-to-Enter	Significance F-to-Remove	Wilk's Lambda
Step 0	(Variables not in)				
	PREDOR	1.0	0.0000		0.27907*
	DORS	1.0	0.0000		0.30123
	HEAD	1.0	0.0000		0.31703
	POSTOR	1.0	0.0000		0.43965
	LFTPECT	1.0	0.0000		0.55968
Step 1	(Variables in)				
	PREDOR	1.0		0.0000	
(Variables not in)					
	DORS	0.9984749	0.0000		0.16388*
	HEAD	0.9134243	0.0000		0.21428
	POSTOR	0.9417196	0.0000		0.24002
	LFTPECT	0.9992086	0.0000		0.22457
Step 2	(Variables in)				
	PREDOR	0.9984749		0.0000	0.30123
	DORS	0.9984749		0.0000	0.27907
(Variables not in)					
	HEAD	0.8688658	0.0000		0.12287*
	POSTOR	0.9387055	0.0000		0.14636
	LFTPECT	0.9987843	0.0000		0.14216
Step 3	(Variables in)				
	PREDOR	0.9126915		0.0000	0.14808
	DORS	0.9497675		0.0000	0.21428
	HEAD	0.8688658		0.0000	0.16388
(Variables not in)					
	POSTOR	0.4506181	0.0000		0.12194
	LFTPECT	0.9140920	0.0000		0.11809*
Step 4	(Variables in)				
	PREDOR	0.9001297		0.0000	0.14408
	DORS	0.9477433		0.0000	0.19597
	HEAD	0.7951900		0.0000	0.14216
	LFTPECT	0.9140920		0.0000	0.12289
(Variables not in)					
	POSTOR	0.4505364	0.0018	0.0000	0.11727*
Step 5	(Variables in)				
	PREDOR	0.8991120		0.0000	0.14324
	DORS	0.9267317		0.0000	0.19593
	HEAD	0.3977323		0.0000	0.13353
	POSTOR	0.4505364		0.0018	0.11809
	LFTPECT	0.9139263		0.0000	0.12194

@minimum tolerance level = 0.001
minimum significance of F-to-enter = 0.01
minimum significance of F-to-remove = 0.01

NOTE: F-values were not included in the above table because of space restrictions.

3.3.3 Discriminant Functions

The following tables (3.2 - 3.7) show Fisher's Linear Discriminant Functions for each of the six stepwise discriminant analysis. All five variables entered and remained in the stepwise procedure for each analysis. Table 3.6 and 3.7 show classification function coefficients for the five common rivers in North America from the 1968 and 1969 data. This allows one to determine the amount of variation in the functions between 1968 and 1969. A better comparison may be obtained when the first two canonical functions are graphed later in Section 3.5. Fisher's linear discriminant functions for the stepwise discrminant analysis of Europe (1969) and North America (1969) are:

$$f_{(\text{Europe})} = -1078.023 + 17.84350(Y_1) + 14.05400(Y_2) + 22.92478(Y_3) - 7.369561(Y_4) + 9.887239(Y_5)$$

$$f_{(\text{N.America})} = -1353.518 + 19.56596(Y_1) + 16.47517(Y_2) + 26.00119(Y_2) - 8.366054(Y_4) + 10.82042(Y_5)$$

The coefficients of the two classification functions are presented in tabular form in Table 3.2. Note that for remaining comparisons, the discriminating functions will be presented in tabular form only.

Table 3.2 Discrimination Coefficients for Europe (1969), North America (1969)

	Europe (1969)	N. America (1969)
PREDOR	17.84350	19.56596
DORS	14.05400	16.47517
HEAD	22.92478	26.00119
POSTOR	-7.369561	-8.366054
LFTPECT	9.887239	10.82042
(constant)	-1078.023	-1353.518

Table 3.3 Discrimination Coefficients for the Five European (1969) Rivers - Logan R., R. Almond, R. Boyne, R. Lee, R. Usk.

	Logan	Almond	Boyne	Lee	USK
PREDOR	32.13552	23.94184	25.60663	28.55470	31.16748
DORS	15.96818	12.17664	13.51858	16.00958	17.28250
HEAD	44.49765	36.00499	35.75827	40.00245	44.13137
POSTOR	-33.94362	-26.94434	-25.22231	-28.93728	-32.76213
LFTPECT	3.926077	5.922640	4.653036	6.172821	5.730297
(constant)	-1696.883	-1062.011	-1162.793	-1487.766	-1734.043

Table 3.4 Discrimination Coefficients for the Six North American (1968) Rivers - Maine, Miramichi, Saint John, Indian R., Salmon R. Salmonier R.

	Maine	Miramichi	Saint John	Indian	Salmon	Salmonier
PREDOR	36.24020	27.80346	32.27585	32.25350	32.53075	30.40244
DORS	31.85607	25.87786	30.01773	28.19062	30.14729	28.64447
HEAD	36.10443	32.54166	35.00321	30.91662	39.41300	37.63644
POSTOR	-1.542365	2.615035	1.726538	6.974666	-0.7203537	-2.800655
LFTPECT	16.62739	14.59497	15.40842	15.03125	16.69534	15.23686
(constant)	-2579.921	-1790.238	-2256.800	-2134.665	-2418.564	-2112.504

Table 3.5 Discrimination Coefficients for the Eight North American (1969) Rivers - Maine, Miramichi, Saint John, Koksoak R., Indian R., Salmon R., Harry's R., Sand Hill R.

	Maine	Miramichi	Saint John	Koksoak
PREDOR	23.87917	19.21030	24.05727	22.25444
DORS	17.53543	15.08423	18.24702	16.68330
HEAD	39.46706	34.84530	37.56235	35.38183
POSTUR	0.7477219	1.638135	6.871874	0.3208730
LFTPECT	10.86295	10.20434	10.52034	11.18892
(constant)	-1933.770	-1425.540	-2009.915	-1666.305

Table 3.5 (cont'd)

	Indian	Salmon	Harry's	Sand Hill
PREDOR	21.41758	20.89572	18.76904	22.76049
DORS	16.23586	15.96027	13.90223	17.81574
HEAD	37.43541	37.83026	32.82964	39.88664
POSTOR	1.004788	-0.3519694	1.047672	2.163779
LFTPECT	9.197749	10.27874	9.726439	10.79514
(constant)	-1631.720	-1610.415	-1285.565	-1913.869

Table 3.6 Discrimination Coefficients for the Five "Common" North American (1968) Rivers - Maine, Miramichi, Saint John, Indian R., Salmon R.

	Maine	Miramichi	Saint John	Indian	Salmon
PREDOR	33.08410	25.35785	29.43074	29.37118	29.71316
DORS	28.70610	23.20265	27.12595	25.31415	27.02385
HEAD	39.23774	35.16950	37.81103	33.87021	42.47194
POSTOR	-6.937315	-1.983088	-3.321890	1.840792	-5.992241
LFTPECT	16.17327	14.17866	14.98599	14.56708	16.23285
(constant)	-2424.869	-1689.012	-2122.323	-2003.848	-2284.901

Table 3.7 Discrimination Coefficients for the Five "Common" North American (1969) Rivers - Maine, Miramichi, Saint John, Indian R., Salmon R.

	Maine	Miramichi	Saint John	Indian	Salmon
PREDOR	29.07089	23.60305	29.45414	26.03939	25.50514
DORS	20.67746	17.76342	21.41219	19.08929	18.81403
HEAD	33.57229	29.96985	31.20832	32.20832	32.73799
POSTOR	4.416097	4.617626	11.18270	4.254854	2.682117
LFTPECT	12.66405	11.44626	12.40768	10.83586	11.71958
(constant)	-2104.985	-1543.269	-2193.042	-1770.514	-1744.209

3.4 Results of Misclassification Probabilities

Recall from Section 3.2 that Hotellings' T^2 test showed that the populations (for example, North America and European rivers) are well seperated. However, this does not mean total non- overlapping of the distributions. Consequently, there remains the possibility that a random observation, may be misclassed into the wrong population. In order to judge the efficiency of the discrimination criterion discussed in the last section, the following procedure is taken: (1) An observation is taken from the existing samples and the discrimina- tion criterion is applied to determine the population in which it belongs. This is repeated and continued for all observations. Next, the total number of cases that were correctly classified, denoted by n_c, is counted and divided by the total number of cases in the sample, denoted by $n_.$. The result is multiplied by 100 to give the percentage of correctly classified cases, denoted by P. Hence, P is calculated by:

$$P = \frac{n_c}{n_.} (100)$$

(2) A proportional reduction in error statistic (Klecker, 1980) gives a standardized measure of improvement regardless of the number of groups. This statistic, called tau, is simply:

$$\text{tau} = \frac{n_c - \sum_{i=1}^{g} p_i n_i}{n_. - \sum_{i=1}^{g} p_i n_i}$$

where n_c and $n_.$ are defined above, n_i is the number of cases in the ith group and p_i is the prior probability of group membership in the ith group. The maximum value for tau is 1.0, and it occurs when there are no errors in prediction. A value of zero indicates no improvement and negative results indicate no discrimination between the groups.

3.4.1 Results of Classification

The following tables contain classification results for each of the six sets of classification functions. Tables 3.8-3.13 give the number of observations in each group (n_i); the number and percentage of observations correctly and incorrectly classified for each group; the percentage of all observations correctly classified, P_i; and the proportional reduction in error statistic, tau.

Table 3.8 gives the classification results for the discriminant analysis between North American and European salmon sampled in 1969. The overall misclassification rate (or error rate) was only 0.35% and a greater but insignificant proportion of European salmon was classified as North American than the converse. The actual proportion of North American

to European-origin salmon was 0.351:0.649, and the predicted proportion from classification was 0.349:0.651. Thus, there is an error rate of 0.2% in favour of European salmon, which is an extremely small percentage.

Table 3.8

European (1969), North America (1969)

Actual Group	n_i	Predicted Group	
		Europe	North America
Europe	495	491 (99.2%)	4 (0.8%)
North America	915	1 (0.1%)	914 (99.9%)

n_i = 1410 P = 99.65%

n_c = 1405 tau = 0.9929

The tau value of 0.9929 indicates that classification based on the five discriminating variables made 99.29% fewer errors than would be expected by random assignment.

Table 3.9 gives the classification results for the discriminant analysis between the salmon sampled from the five European rivers in 1969. The overall misclassification rate was 3.84%, ranging from 1.3% for River Almond to 8.3% for River Usk. The tau value of 0.9520 indicates that classification based on the five discriminating variables made 95.20% fewer errors than would be expected by random assignment.

Table 3.10 displays the classification matrix for the discriminant

analysis between the salmon sampled from the six North American rivers in 1968. The overall misclassification rate was only 5.52%, ranging from 0.0% for Miramichi to 13.5% for Saint John. The tau value of 0.9337 shows that classification based on the five discriminating variables made 93.87% fewer errors than would be expected by random assignment.

In Table 3.11, the classification results for the discriminant analysis between the salmon sampled from each of the eight North American rivers in 1969 is given. The overall misclassification rate for the eight rivers was only 17.13%, ranging from 5.6% for Harry's River to 33.8% for Sand Hill River. The tau value of 0.8042 shows that classification based on the five discriminating variables made 80.42% fewer errors than would be expected by random assignment.

Tables 3.12 and 3.13 give the classification results for the discriminant analysis between the salmon sampled from each of the five "common" rivers in 1968 (Table 3.12) and 1969 (table 3.13). The overall misclassification rate for 1968 was 4.54% whereas it was 11.47% for the 1969 data. That is, more than $2\frac{1}{2}$ times more salmon were misclassified for the same regions in 1969. In 1968, the range of misclassification went from 0.0% for Miramichi to 8.1% for Saint John. In 1969, the range was from 4.5% to 26.8% for Miramichi and Salmon River respectively. The tau values of 0.9433 for 1968 and 0.8567 for 1969 indicate that the classification based on the discriminating variables made 94.33% and 85.67% fewer errors respectively than would be expected by random assignment.

Table 3.9

European Rivers

Actual Group	n_i	Actual Group				
		Logan	Almond	Bayne	Lee	Usk
Logan	117	110 (94.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (6.0%)
Almond	158	0 (0.0%)	156 (98.7%)	2 (1.3%)	0 (0.0%)	0 (0.0%)
Boyne	50	0 (0.0%)	0 (0.0%)	49 (98.0%)	1 (2.0%)	0 (0.0%)
Lee	98	0 (0.0%)	0 (0.0%)	0 (0.0%)	95 (96.9%)	3 (3.1%)
Usk	72	6 (8.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	66 (91.7%)

$n_. = 495$ P = 96.16%

$n_c = 476$ tau = 0.9520

Table 3.10

North American (1968) Rivers

Actual Group	n_i	Predicted Group					
		Maine	Miramichi	Saint John	Indian	Salmon	Salmonier
Maine	81	78 (96.5%)	0 (0.0%)	1 (1.2%)	0 (0.0%)	2 (2.5%)	0 (0.0%)
Miramichi	147	0 (0.0%)	147 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Saint John	148	1 (0.7%)	0 (0.0%)	128 (86.5%)	3 (2.0%)	8 (5.4%)	8 (5.4%)
Indian	50	0 (0.0%)	0 (0.0%)	2 (4.0%)	48 (96.0%)	0 (0.0%)	0 (0.0%)
Salmon	147	0 (0.0%)	0 (0.0%)	7 (4.8%)	0 (0.0%)	139 (94.6%)	1 (0.7%)
Salmonier	151	0 (0.0%)	0 (0.0%)	7 (4.6%)	0 (0.0%)	0 (0.0%)	144 (95.4%)

$n_i = 724$ P = 94.48%

$n_c = 684$ tau = 0.9337

Table 3.11

North American (1969) Rivers

Actual Group	n_i	Predicted Group							
		Maine	Mira-michi	Saint John	Koksoak	Indian	Salmon	Harry's	Sand Hill
Maine	142	119 (83.8%)	0 (0.0%)	4 (2.8%)	0 (0.0%)	1 (0.7%)	0 (0.0%)	0 (0.0%)	18 (12.7%
Miramichi	151	0 (0.0%)	139 (92.1%)	0 (0.0%)	1 (0.7%)	0 (0.0%)	3 (2.0%)	8 (5.3%)	0 (0.0%
Saint John	73	4 (5.5%)	0 (0.0%)	62 (84.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (9.6%
Koksoak	130	0 (0.0%)	0 (0.0%)	0 (0.0%)	117 (90.0%)	5 (3.8%)	6 (4.6%)	0 (0.0%)	2 (1.5%
Indian	125	0 (0.0%)	2 (1.6%)	0 (0.0%)	5 (4.0%)	95 (76.0%)	23 (18.4%)	0 (0.0%)	0 (0.0%
Salmon	41	0 (0.0%)	1 (2.4%)	0 (0.0%)	0 (0.0%)	9 (22.0%)	31 (75.6%)	0 (0.0%)	0 (0.0%
Harry's	89	0 (0.0%)	5 (5.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	84 (94.4%)	0 (0.0%
Sand Hill	148	26 (17.6%)	0 (0.0%)	23 (15.5%)	1 (0.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	98 (66/2%

n_i = 899 P = 82.87%

n_c = 745 tau = 0.8092

Table 3.12

Common North American (1968) Rivers

Actual Group	n_i	Predicted Group				
		Maine	Miramichi	Saint John	Indian	Salmon
Maine	81	77 (95.1%)	0 (0.0%)	1 (1.2%)	0 (0.0%)	3 (3.7%)
Miramichi	147	0 (0.0%)	147 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Saint John	148	1 (0.7%)	0 (0.0%)	136 (91.9%)	3 (2.0%)	8 (5.4%)
Indian	50	0 (0.0%)	0 (0.0%)	2 (4.0%)	48 (96.0%)	0 (0.0%)
Salmon	147	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	139 (94.6%)

$n_.$ = 573 p = 95.46%

n_i = 547 tau = 0.9433

Table 3.13

Common North American (1969) Rivers

Actual Group

	Maine	Miramichi	Saint John	Indian	Salmon
Maine	133 (93.7%)	0 (0.0%)	7 (4.9%)	1 (0.7%)	1 (0.7%)
Miramichi	0 (0.0%)	144 (95.4%)	0 (0.0%)	0 (0.0%)	7 (4.6%)
Saint John	7 (9.6%)	0 (0.0%)	66 (90.4%)	0 (0.0%)	0 (0.0%)
Indian	0 (0.0%)	2 (1.6%)	0 (0.0%)	98 (78.4%)	25 (20.0%)
Salmon	0 (0.0%)	2 (4.9%)	0 (0.0%)	9 (22.0%)	30 (73.2%)

$n_. = 532$ P = 88.53%

$n_i = 471$ tau = 0.8567

Notice here that the observations used to determine the discriminating functions were also used to calculate the percentage of cases correctly classified. Many authors (c.f. Lachenbruch and Mickey, 1968; Srivastava and Carter, 1983) suggest that this method of estimating classification rate tends to overestimate the power of the classification procedure because the validation is based on the same cases used to derive the classification functions. The next section will determine the reality of this problem for the analysis.

3.5 Verification of Classification Results: The Jackknife Technique

The classification procedure used in the previous section is verified by using the jackknife classification technique. This technique is used to remove some of the bias inherent in basing classification decisions

on that data set used to determine the classification functions. However, it has been determined that when large samples are available, it is not necessary to use the jackknifing technique because the bias has already been reduced to a very low level. This will be shown in the following subsection.

3.5.1 Generation of Random Numbers to Select Observations

For the jackkifing technique, random numbers are generated such that each observation has a predetermined probability of being selected. For example, given that the total sample size is n and one wants, on the average, h specimens excluded for each jackknife, then one would choose

$$p = 1 - \frac{h}{n} .$$

To determine which of the n specimens are excluded, a number ranging from 0 to 1 is assigned to each of the n observations. If the random number for a particular specimen is p or less, then that observation remains. However, if the random number is greater than p, then that observation is excluded. The observations not excluded are then used to determine the classification functions and these functions are used to classify the remaining unselected observations.

3.5.2 Results

(A) North America (1969) versus Europe

For the discriminant analysis of North American and European origin salmon, there is a total of 1410 observations. Using the jackknife technique (Appendix B), samples were taken such that each observation had a probability of 0.9858 of being selected. This gives approximately 1390 selected observations and 20 unselected observations per sample (refer to Appendix B). The discriminating coefficients were then determined for each sample and were used to classify the remaining unselected observations. One can now calculate the number classified correctly and incorrectly for these unselected cases. This was repeated 500 times for a total of 9991 unselected observations classified. The results showed 32 misclassified and 9959 correctly classified cases, i.e. 0.32% misclassified and 99.68% correctly classified (see last row of the table in Appendix B). This percentage is actually 0.03% higher than the 99.65% originally classified correctly. Therefore, no bias was evident in this analysis.

(B) Five European Rivers

This jackknife procedure was used on the European origin salmon taken from five rivers in 1969. Samples were taken from the 495 observations such that each observation has a probability of 0.9494 of being selected. This gives approximately 470 selected and 25 unselected cases per sample. The unselected cases were again classified by using the discriminant functions derived from the selected observations. This was repeated 400 times for a total of 10,062 unselected cases of which

470 observations (4.67%) were misclassified and 9592 observations (95.33%) were correctly classified (Appendix C). This classification rate is only 0.83% lower than the original rate of 96.16%.

In summary, it was found that for $n = 495$, the bias was less than 1% and for $n = 1410$, the bias was negligible. Therefore, if one assumes a decreasing bias for an increasing sample size, then it is safe to base classification decisions on the data sets used to determine the classification functions. That is, the samples are large enough to reduce the bias to a minimal level.

3.6 Canonical Discriminant Functions

Classification can also be done with the canonical variables instead of using the original discriminating variables. The final classifications will generally be identical; however, a better picture of how cases are being classified can be obtained by superimposing the classification boundary lines over a plot of cases. These classification plots are useful for examining the relationship of groups to each other and graphically depicting misclassifications. In general, the first n canonical variables will produce an n dimensional graph. Therefore, for convenience, the first two canonical variables will be plotted. The underlying theory for this methodology is explained in the following:

Suppose there are p variables in a discriminant analysis of g groups. It is desirable to find new variables that are independent and have the largest F-values for testing equality of the g means. Thus,

one wishes to find a vector

$\underset{\sim}{a} = (a_1, \ldots, a_p)'$ such that

$$\sum_{i=1}^{g} n_i (\underset{\sim}{a}'\bar{y}_i - \underset{\sim}{a}'\bar{y})^2 (\underset{\sim}{a}'\underset{\sim}{S}_p\underset{\sim}{a})^{-1} \qquad [3.6.1]$$

is a maximum (cf. Srivastava and Carter, 1983), where $\bar{y}_i$ is the mean vector of the ith population, n_i is the number of observations from the ith population, $\bar{y}$ is the average of all the observations, and $\underset{\sim}{S}_p$ is the pooled covariance matrix (see Section 3.2).

The maximum of 3.6.1 occurs when $\underset{\sim}{a}$ satisfies the equation

$$(B - \lambda \underset{\sim}{S}_p)\underset{\sim}{a} = 0 \qquad [3.6.2]$$

where B is the between groups mean sum of squares given by

$$\frac{1}{g-1} \sum_{i=1}^{g} n_i (y_i - \bar{y})(y_i - \bar{y})' ,$$

and λ is the maximum eigenvalue of $\underset{\sim}{S}_p^{-1}B$.

Since the first two canonical variables are of interest (as long as the minimum of p and $g - 1$ is greater than one), one calculate

$$t_i = \underset{\sim}{a}_i'y \qquad i = 1, 2 ,$$

where a_i is the solution of 3.6.2 for λ equal to the ith largest eigenvalue of $S^{-1}B$. Then, the first two canonical variables are plotted to show the separation of the g groups.

To get a better picture of how cases are being classified,

the classification boundary lines are superimposed over the plot of cases. In Figures 3.1 to 3.4, the broken lines seperating the groups represent these classification boundaries and the solid lines represent the boundaries for the plot of cases. Note here that a plot for the analysis of European origin salmon versus North American origin salmon was not included because there was only one canonical discriminant function obtained from the two groups. Also, the plot for the analysis of the eight North American rivers sampled in 1968 was not included because with so many plots of cases, it was difficult to distinguish one plot from the other. The results are as follow:

(1) Five European Rivers

Figure 3.1 shows the plots and boundaries of the first two canonical discriminant functions extracted from the stepwise discriminant analysis procedure. These two canonical functions represent 98.69% of the total variability between the groups. River Almond (2) and River Boyne (3) are well seperated from the other rivers. However, Logon River (1) and River Usk (5) have a large percentage of overlap, which leads to a large misclassification between them.

(2) Six North American Rivers - 1968

Figure 3.2 shows the plots and boundaries of the first two extracted canonical discriminant functions. These two canonical functions represent 94.57% of the total variability between the groups. Miramichi (2) is well seperated from all other groups, but Saint John (3) tends to overlap Indian River(4), Salmon

River (5) and Salmonier River (6). Otherwise, the groups of Maine (1), Indian River, Salmon River and Salmonier River are seperated relatively well.

(3) Five Common Rivers of North America (1968)

Figure 3.3 shows the plots and boundaries of the first two extracted canonical discriminant functions. The two functions represent 96.45% of the total variability between groups. The figure indicates Miramichi (2) as having good separation from the other groups. However, some overlap exists between Saint John (3) and Indian River (4); and Saint John and Salmon River (5).

(4) Five Common Rivers of North America (1969)

Figure 3.4 shows the plots and boundaries of the first two extracted canonical discriminant functions. The two functions represent 98.46% of the total variation between groups. Maine (1) and Saint John (3) are clearly seperated from the remaining groups. However, there is overlap present between Maine and Saint John; and Indian River (4) and Salmon River (5) have relatively large overlaps. This large overlap is evident in the classification table (Table 3.13).

Since Figures 3.3 and 3.4 are plots of the same rivers sampled in 1968 and 1969, one would expect some similarity. However, the classification percentages were not the same (95.46% and 88.53%) and the positions of the group plots shifted very significantly. For example, Indian River and Salmon River were virtually without overlap in 1968, however a relatively large overlap existed in 1969.

Also, Miramichi, which overlapped with Indian River and Salmon River in 1968, was very well seperated from these two rivers in 1969. The reasons for these differences are not determined here but is left for further study.

There are certain situations when the classifications and canonical discriminant functions will not necessarily provide the same results. In particular, this is true when the group covariance matrices are not equal. This is because the pooled variance-covariance matrix must be used when calculating the canonical discriminant functions. Unfortunately, there is no clear guidelines for determining how different the group covariance matrices must be before the use of canonical discriminant functions becomes unjustified. However, Tatsuoka (1971, p. 232-33) reports evidence that the canonical discriminant function procedure yields similar results and can be used unless the group covariance matrices are "drastically" different. From this point of view, tests regarding equality of variances might have been more appropriate but were not chosen in the present report.

Figure 3.1 Plots and Boundaries of the first two Canonical Variates for the Five European Rivers.

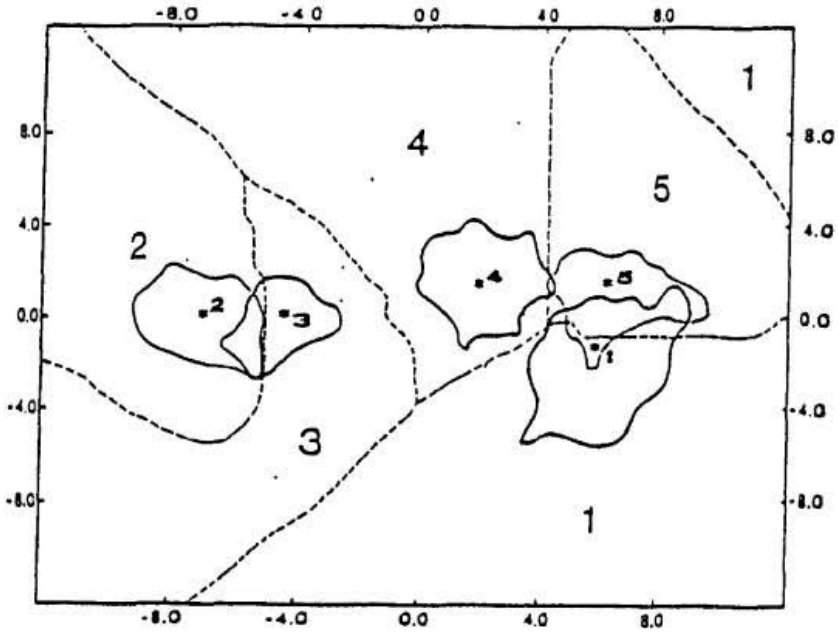


Figure 3.2 Plots and Boundaries of the first two Canonical Variates for the Six North American Rivers (1969)

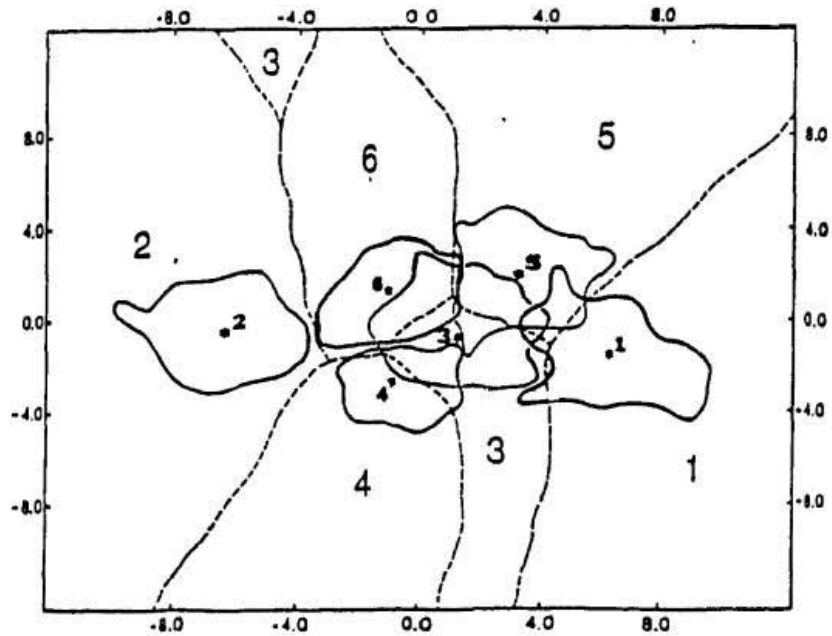


Figure 3.3 Plots and Boundaries of the first two Canonical Variates for the Five Common North American Rivers (1968)

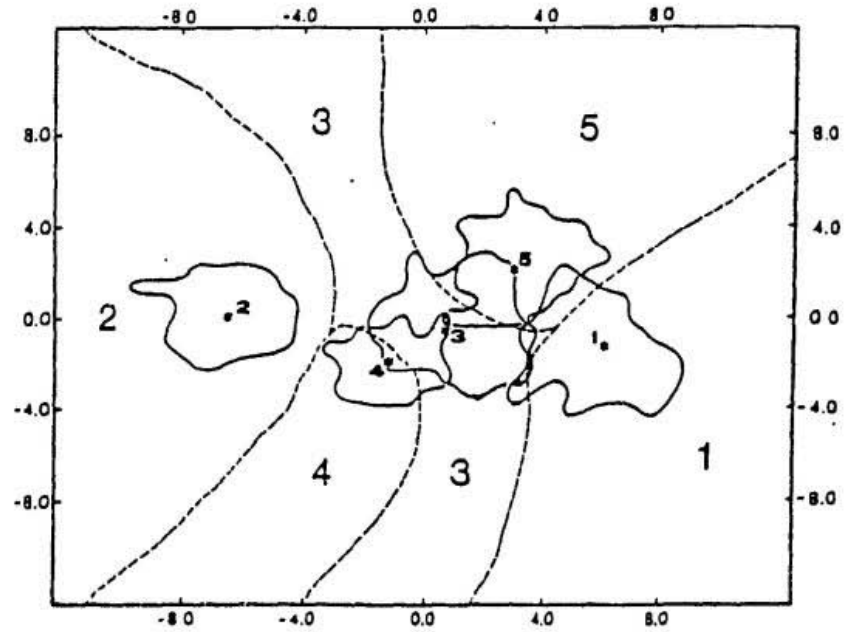
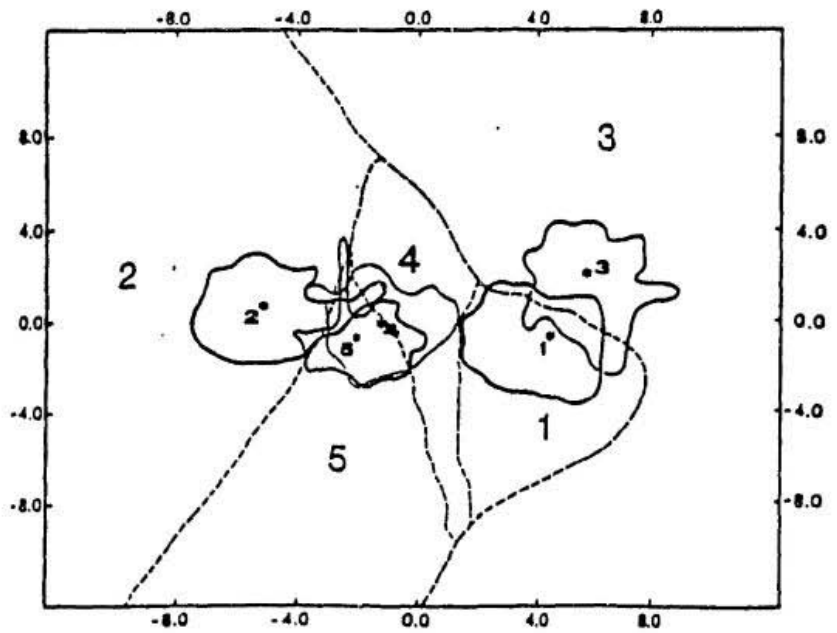


Figure 3.4 Plots and Boundaries of the first two Canonical Variates for the Five Common North American Rivers (1969)



Chapter 4

CLUSTERING APPROACH FOR DISCRIMINATION

4.0 Introduction

It is well known that the basic aim of cluster analysis is to find the "natural groupings", if any, of a set of specimens. Thus, cluster analysis aims to allocate the set of specimens to a set of mutually exclusive, exhaustive groups such that specimens within a group are similar to one another while being dissimilar from specimens in other groups. In discriminant analysis, one begins with apriori well defined groups and asks how the given groups differ. However, in cluster analysis, one begins with a group and asks whether the given group can be partitioned into sub-groups that differ in some meaningful way.

As cluster analysis is, in general, able to construct seperate groups, the techniques of clustering are applied to the adjusted data $(Y_1, Y_2, Y_3, Y_4, Y_5)$ and examine the validity of seperation between groups is examined. Thus, the aim of the present analysis is fundamentally different from that of usual cluster analysis. More specifically, the adjusted data of k groups is combined to form a single data set and then clustering techniques are applied to see whether the k groups are well seperated or not. Hence, it is a discrimination analysis based on

the clustering principle.

One may describe the clustering principle, in general, as follows: First, k measurements are taken on each of the n specimens. The n by k matrix of raw data is then transformed into an n by n matrix of distance measures where the distances are computed between pairs of objects across the k variables. Next, a clustering algorithm is selected, which defines the rules concerning how to cluster the objects into subgroups on the basis of the distance measures. Finally, the uncovered clusters are contrasted, or profiled, in terms of their mean values on the k variables or other characteristics of interest.

4.1 Similarity Measures

Fundamental to the use of any clustering technique is the computation of a measure of similarity or distance between the objects (specimens) concerned. These distance measures can be seperated into two broad classes in two distinct ways, depending on the nature of the data. For data having qualitative components, a matching-type measure is appropriate. However, since the data is quantitative, a distance-type measure will be used.

Each k-dimensional specimen is represented by the vector $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_k)$ where $k = 5$. The notation y_i is used to denote the measurements collected on the ith specimen, that is, $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$. The familiar Euclidean distance, d_{ij}, between two specimens i and j is denoted as

$$d_{ij} = \left[\sum_{\ell=1}^{k} (Y_{i\ell} - Y_{j\ell})^2 \right]^{1/2} , \quad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, n \\ i \neq j \end{matrix} \qquad [4.1.1]$$

and the squared-Euclidean distance is

$$d_{ij}^2 = \sum_{\ell=1}^{k} (Y_{i\ell} - Y_{j\ell})^2 , \quad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, n . \\ i \neq j \end{matrix} \qquad [4.1.2]$$

This will be used as a basis to define appropriate Euclidean distances in Section 4.3 for the purpose of clustering.

4.2 Clustering Techniques

The next step is to select a particular type of computational algorithm. Two of the most popular types of clustering techniques are hierarchical and partitioning. Hierarchical techniques cluster the clusters themselves at various levels, whereas partioning techniques form clusters by optimizing some specific clustering criterion.

Hierarchical Techniques perform successive fusions or divisions of the data. One of the main features distinguishing hierarchical techniques from other clustering algorithms is that once an object joins a cluster, it is never removed and fused with other objects belonging to some other cluster. Agglomerative methods proceed by forming a series of fusions of the n specimens into groups. Divisive methods partition the set of n specimens into finer and finer subdivisions. The output from these methods is typically summarized by the use of a dendrogram. This is a two-dimensional

tree-like diagram illustrating the fusions or partitions that have been constructed at each successive level. Everitt (1980), Dillon and Goldstein (1984), and Chatfield and Collins (1980), among others, discuss these techniques in further detail.

Partitioning Techniques Unlike hierarchical clustering techniques, methods that affect a partition of the data do not require that the allocation of an object be irreversible. Thus, objects may be reassigned if their initial placements are inaccurate. These techniques partition the data based upon optimizing some predefined criterion. The use of partitioning techniques usually assumes that the number of final clusters is known and specified in advance, although some methods will allow the number to vary. There are many partitioning techniques, and they differ with respect to (1) how clusters are initiated, (2) how objects are allocated to clusters, and (3) how some or all of the objects already clustered are reallocated to other clusters.

For the data used in this study, it i felt that partitioning techniques were selected because it was desirable to obtain a predefined number of clusters. In other words, a partitioning algorithm allows one to specify the final number of clusters in advance. The algorithm produces clusters by finding cluster centres based on the values of the cluster variables and assigns cases to the centres that are nearest. The basis of this partitioning algorithm is described in the following section.

4.3 Algorithm for Partitioning Technique

Denote the conditional variable Y_{ij}, the value of the jth specimen on the ith variable, i = 1, 2, ..., k; j = 1, 2, ..., n, as before. Let P_{nk} be the partition that results in each of the n specimens to be allocated to one of g clusters. The mean of the ith variable in the ℓth (ℓ = 1, ..., g) cluster will be denoted by $\bar{Y}_{i\ell}$, and the number of individuals belonging to the ℓth cluster by n_ℓ. Following equation 4.1.2, the squared Euclidean distance between the jth specimen and ℓth cluster is expressed as

$$D_{j\ell} = \sum_{i=1}^{k} (Y_{ij} - \bar{Y}_{i\ell})^2 \qquad [4.3.1]$$

The error component of the partition is defined as

$$E[P_{nk}] = \sum_{j=1}^{n} D_{j,\ell(j)} \qquad [4.3.2]$$

where $\ell(j)$ is the cluster that contains the jth specimen, and $D_{j,\ell(j)}$ is the squared Euclidean distance between specimen j and the cluster mean of the cluster containing the specimen. The procedure is as follows:

(1) Firstly, the initial cluster centres are selected. A centre is an estimate of the average value of each clustering variable for the cases in a cluster. (A centre includes one value for each variable). This can be obtained in various ways. One method is to select the k

cases with well seperated values as initial centres, where k is the number of final clusters desired. Then, the sample means of the variables can be used for each group as the initial cluster centres.

(2) Next, the values of the initial cluster centres are updated to derive the classification cluster centres. Each case is assigned, in turn, to the nearest cluster centre (measured by the squared Euclidean distance, $D_{i\ell}$) such that $E[P_{nk}]$ (equation 4.1.4) is minimized. When a case is assigned, the procedure updates the centre to a mean for the cases that are thus far in the cluster. Therefore, as the cases are processed, the centres migrate to concentrations of observations.

(3) The final step ressigns each case to the nearest of the updated (classification) cluster centres. The reassignment yields the final clusters, and the final cluster centres result from the variable means for the cases in the final cluster.

4.4 Construction of Appropriate Clusters based on Partitioning Techniques

a) The k Most Seperated Observation as Initial Centres

The above procedure, using the k most seperated observations as the initial cluster centres, were used to cluster the samples of (1) North American salmon sample in 1969, (2) North American salmon (1968), (3) European salmon (1969) and (4) the combined sample of North American and European salmon sampled in 1969. Since the number of groups are known for each of the four samples, k is initialized to

equal the number of groups represented in the sample.

However, the results of this method were flawed by the presence of extreme values in some of the groups. Although outliers were removed from the data (re: Chapter 2), there were relatively extreme cases that remained. As a result, some of these cases were chosen as initial cluster centres and, since these cases are far removed from the rest of the data, no observations (or very few) were assigned to them. Thus, it resulted in some clusters containing few observations (sometimes only one), while other clusters contained a large portion of the data.

b) The k Sample Means as Initial Cluster Centres

A much more effective method can be used by taking the k sample means as the initial cluster centres. The procedure of Section 4.3 was again implemented using these k sample means instead of k most seperated observations. This method was used to cluster the samples of (1) North American (1969) salmon, (2) North American (1968) salmon, (3) European salmon, and (4) North American (1969) and European salmon. Again, k is utilized to equal the number of groups represented in the sample. The following section describes the results of this clustering procedure applied to the four populations.

1. The Eight Regions of North America (1969)

The above procedure was used to cluster the sample of North American (1969) salmon into eight groups. Eight cluster centres were initialized (Table 4.1) from the eight vector means. Tables 4.2 and

4.3 show the classification and final cluster centres respectively, which were calculated on the basis of equations 4.3.1 and 4.3.2. Table 4.4 display the squared Euclidean distances, d_{ij}^2 (equation 4.1.2), between all pairwise final cluster centres i and j.

Since the origin of the observations are already known, a classification table can be produced to determine the results of the clustering procedure (Table 4 5). In a practical situation, a classification table would not be used when doing a clustering procedure. This is because a cluster analysis is usually only used when the origin of the specimens is not known. However, it is very informative for the purpose of this study.

It is shown in the classification table that the clustering procedure maintained 72.64% of the original groupings. This percentage suggests that there is good seperation between the eight groups. Recall that using discriminant analysis for this sample, the percentage of correctly classified cases was 82.87%. This difference of approximately 10% is not unlikely. Since the clustering procedure does not take into consideration the variance-covariance matrix, a lower classification rate is expected using this method.

2. The Six Regions of North America (1968)

This procedure was again used to cluster the sample of North American (1968) salmon into six groups. Six cluster centres were initialized (Table 4.6) from the six vector means. Tables 4.7 and 4.8 show the classification and final cluster centres respectively

and Table 4.9 displays the Euclidean distances between all pairwise final cluster centres.

The results of the classification table indicates that the clustering procedure maintained 86.05% of the original groupings (Table 4.10). This suggests that the six groups are well seperated. Notice that this classification table is very similar to Table 3.10 of Chapter 3 according to where groups are misclassified.

3. The Five Rivers of Europe

The third analysis deals with the clustering of the sample of European salmon into five groups. Five cluster centres were once again initialized (Table 4.11) from the five vector means. The classification and final cluster centres are displayed in Tables 4.12 and 4.13 respectively and the Euclidean distances between all pairwise final cluster centres are given in Table 4.14.

Table 4.15 displays the classification results for the clustering procedure. The percentage of correctly grouped observations is 85.25%. This suggests that the group meansare well seperated. Other than River Usk, this table is very similar to the classification table for these rivers in Chapter 3 (Table 3.10).

4. The Two Groups of North America (1969) and Europe

Finally, the clustering procedure was used to cluster the sample of North American (1969) and European origin salmon into two groups.

The two cluster centres were initialized (Table 4.16) from the two vector means. Tables 4.17 and 4.18 show the classification and final cluster centres respectively and Table 4.19 displays the Euclidean distance between the two cluster centres.

The results of the classification table indicate that the clustering procedure maintains 99.57% of the original groupings (Table 4.20). This is an extremely high percentage, and is only slightly less than the classification percentage of the classification table of Chapter 3 (Table 3.8). This is a strong indication that these two groups are very well seperated.

In summary, all clustering methods maintain a high percentage of the original groupings. Compared to the discriminant analysis, the clustering procedure failed to seperate approximately 10% more salmon than the discriminant analysis. This is because, as explained earlier, the variance-covariance matrix is not used in clustering procedures. However, the clustering did determine the seperation amongst groups and supported the results of the Hotelling's T^2 statistics and tests of Chapter 3.

Table 4.1 Initial Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	63.110	35.199	36.101	18.570	27.657
2	52.582	29.852	31.713	16.473	24.940
3	63.904	36.937	36.636	19.940	27.677
4	58.402	34.057	32.704	16.822	26.639
5	57.398	32.350	33.665	17.330	24.686
6	56.529	31.595	33.622	17.080	25.807
7	50.693	27.374	30.093	15.542	23.720
8	61.501	35.648	36.336	18.924	27.643

Table 4.2 Classification Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	64.0166	34.7222	36.6875	19.1648	27.5777
2	52.8225	30.7478	31.7491	16.3737	24.8915
3	63.7125	37.0862	36.5809	19.9317	27.8444
4	59.0299	34.0170	33.4704	17.2110	26.6839
5	57.3588	32.3291	33.8912	17.3433	24.6934
6	56.3060	31.3122	33.4865	17.0123	25.8056
7	50.9384	27.3986	30.0396	15.4655	23.7447
8	61.1235	36.3412	36.0504	18.9852	27.4767

Table 4.3 Final Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	63.5578	34.5053	36.4445	18.8752	27.7362
2	52.5769	30.1946	31.7340	16.4852	24.9693
3	63.8963	37.2914	36.5865	19.5530	28.0397
4	58.6547	34.3405	32.9565	16.9651	26.6416
5	57.6525	32.4228	33.6967	17.3173	24.4655
6	56.4772	31.3439	33.3148	17.0795	25.9564
7	50.8619	27.4548	30.3050	15.6610	23.8427
8	61.0525	35.9335	35.9784	18.6947	27.3939

Table 4.4 Distances Between Final Cluster Centres

CLUSTER	1	2	3	4	5	6	7	8
1	0.0000							
2	13.2181	0.0000						
3	2.9067	14.8620	0.0000					
4	6.4094	7.6583	7.6165	0.0000				
5	7.7384	5.9603	9.4240	3.1764	0.0000			
6	8.7359	4.5122	10.5631	3.7858	2.2295	0.0000		
7	16.5553	3.7997	18.4085	11.1671	9.2428	7.8864	0.0000	
8	2.9468	11.5562	3.3845	4.5800	6.2899	7.3326	15.1571	0.0000

Table 4.5 Classification Results for Cluster Analysis of the Eight North American (1969) Rivers

Actual Group	Cluster 1	2	3	4	5	6	7	8	n_i
Maine	83 (58.5%)	0 (0.0%)	26 (18.3%)	2 (1.4%)	1 (0.7%)	0 (0.0%)	0 (0.0%)	30 (21.1%)	14
Miramichi	0 (0.0%)	121 (80.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (6.0%)	21 (13.9%)	0 (0.0%)	15
Saint John	16 (21.9%)	0 (0.0%)	48 (65.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (12.3%)	7
Koksoak	0 (0.0%)	0 (0.0%)	0 (0.0%)	101 (77.7%)	6 (4.6%)	19 (14.6%)	0 (0.0%)	4 (3.1%)	13
Indian	0 (0.0%)	2 (1.6%)	0 (0.0%)	9 (7.2%)	90 (72.0%)	24 (19.2%)	0 (0.0%)	0 (0.0%)	12
Salmon	0 (0.0%)	3 (7.3%)	0 (0.0%)	1 (2.4%)	8 (19.5%)	29 (70.7%)	0 (0.0%)	0 (0.0%)	4
Harry's	0 (0.0%)	2 (2.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	87 (97.8%)	0 (0.0%)	8
Sand Hill	22 (14.9%)	0 (0.0%)	24 (16.2%)	8 (5.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	94 (63.5%)	14

$n_. = 899$ P = 72.64%

$n_c = 653$ tau = .6873

Table 4.6 Initial Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	65.020	36.207	34.326	15.985	25.972
2	51.989	28.849	30.931	15.090	23.220
3	58.968	34.449	33.257	16.070	24.517
4	58.820	31.608	32.360	16.602	24.036
5	60.702	33.552	35.891	16.805	26.647
6	56.501	32.210	33.266	15.189	24.383

Table 4.7 Classification Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	64.0886	35.8620	34.4046	15.9950	25.8256
2	52.0625	28.8580	30.7895	15.1945	23.1692
3	59.1832	35.1958	33.2843	16.0499	24.3491
4	58.9611	32.1427	32.6099	16.2438	24.1071
5	60.8985	33.1519	35.4816	16.8014	26.4011
6	56.8950	32.4651	33.3535	15.5177	24.4467

Table 4.8 Final Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	64.9837	36.1880	34.3453	16.0108	25.9506
2	51.9893	28.8470	30.9314	15.0898	23.2198
3	59.2109	35.1718	33.3149	16.0349	24.5755
4	59.0975	31.8416	32.7960	16.2792	24.1094
5	60.8492	38.3023	36.1153	16.9605	26 8929
6	56.4882	32.3731	33.2206	15.3088	24.4011

Table 4.9 Distances Between Final Cluster Centres

CLUSTER	1	2	3	4	5	6
1	0.0000					
2	15.5790	0.0000				
3	6.1083	10.0283	0.0000			
4	7.7071	8.0729	3.4130	0.0000		
5	5.5085	11.9253	4.4998	4.9431	0.0000	
6	9.5334	6.2736	3.9764	2.8799	6.0990	0.0000

Table 4.10 Classification Results for Cluster Analysis of the Six North American (1968) Rivers (Regions)

Actual Group	Cluster 1	2	3	4	5	6	n_i
Maine	79 (97.5%)	0 (0.0%)	1 (1.23%)	0 (0.0%)	1 (1.23%)	0 (0.0%)	81
Miramichi	0 (0.0%)	147 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	147
Saint John	1 (0.7%)	0 (0.0%)	95 (64.2%)	24 (16.2%)	9 (6.1%)	19 (12.8%)	148
Indian	0 (0.0%)	0 (0.0%)	1 (2.0%)	41 (82.0%)	2 (2.0%)	6 (12.0%)	50
Salmon	3 (2.0%)	0 (0.0%)	18 (12.2%)	5 (3.4%)	119 (81.0%)	2 (1.4%)	147
Salmonier	0 (0.0%)	0 (0.0%)	3 (2.0%)	6 (4.0%)	0 (0.0%)	142 (94.0%)	151

$n_. = 724$ P = 86.05%

$n_c = 623$ tau = .8326

Table 4.11 Initial Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	59.731	30.625	32.922	16.961	23.635
2	45.564	22.391	27.802	14.052	22.775
3	48.730	25.688	28.166	14.980	21.436
4	54.454	30.515	31.329	16.538	24.980
5	58.971	33.025	33.525	17.442	25.840

Table 4.12 Classification Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	59.4858	30.4303	32.7975	16.7572	23.2396
2	45.7243	22.3127	27.8574	14.1128	22.5694
3	47.7626	25.2664	27.9760	14.7471	21.6014
4	56.3029	32.0214	32.1005	16.7309	25.3049
5	60.8637	31.9429	33.9937	17.7272	25.5635

Table 4.13 Final Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	59.3660	30.2502	32.6874	16.7580	23.2617
2	45.5338	22.2952	27.8050	14.0449	22.7869
3	48.3848	25.4977	28.0220	14.8763	21.5343
4	55.2552	31.1751	31.7301	16.6399	25.0713
5	60.4339	32.5288	33.9062	17.7790	25.6889

Table 4.14 Distances Between Final Cluster Centres

CLUSTER	1	2	3	4	5
1	0.0000				
2	16.9126	0.0000			
3	13.0944	4.5488	0.0000		
4	4.6860	14.1674	10.4310	0.0000	
5	3.8408	19.6552	15.9665	5.9217	0.0000

Table 4.15 Classification Results for Cluster Analysis of the Five European Rivers

Actual Group	1	2	3	4	5	n_i
Logan	85	0	0	6	26	117
	(72.6%)	(0.0%)	(0.0%)	(5.1%)	(22.2%)	
Almond	0	152	6	0	0	158
	(0.0%)	(96.2%)	(3.8%)	(0.0%)	(0.0%)	
Boyne	0	0	49	1	0	50
	(0.0%)	(0.0%)	(98.0%)	(2.0%)	(0.0%)	
Lee	1	0	0	97	0	98
	(1.0%)	(0.0%)	(0.0%)	(99.0%)	(0.0%)	
Usk	5	0	0	28	39	72
	(6.9%)	(0.0%)	(0.0%)	(38.9%)	(54.2%)	

$n_. = 495$ P = 85.25%

$n_i = 422$ tau = .8157

Table 4.16 Initial Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	52.942	27.825	30.575	15.818	23.725
2	58.306	33.125	33.954	17.620	26.216

Table 4.17 Classification Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	54.0904	27.4278	30.7977	15.7779	23.1568
2	58.3763	33.0584	33.6789	17.7684	25.9015

Table 4.18 Final Cluster Centres

CLUSTER	PREDORSAL	DORSAL	HEAD	POSTORBITAL	LEFT PECTORAL
1	52.9039	27.7859	30.5753	15.8166	23.7175
2	58.3034	33.1229	33.9385	17.6132	26.2097

Table 4.19 Distances Between Final Cluster Centres

CLUSTER	1	2
1	0.0000	
2	8.8537	0.0000

Table 4.20 Classification Results for Cluster Analysis of North American (1969) and European (1969) Rivers

Actual Group	Cluster 1	Cluster 2	n_i
Europe	490 (99.0%)	5 (1.0%)	495
North America	1 (0.1%)	914 (99.9%)	915

$n_. = 1410$ P = 99.57%

$n_c = 1404$ tau = 1404

CHAPTER 5

CONCLUSIONS

Initially, the underlying distribution of the data was determined in Chapter 2 by means of "Exploratory Data Analysis". By the use of graphical and numerical summaries, normality was indicated for each of the three data groups. Next, the method of shifted power transformation was used to confirm that insignificant departure from multivariate normality existed and that no transformation of the data was necessary for any of the three data groups.

Analysis of covariance was applied to adjust each of the variables for each group to the overall mean standard length in Section 3.2 of Chapter 3. By applying the Mahalanobis generalized sample squared distance technique to the adjusted variables, it ws found that the populations were significantly different pairwis.'.

A quadratic stepwise discriminant analysis (Section 3.3, Chapter 3) gave the best results using Fisher's linear discriminant functions. For each of the six analyses, all five conditional variables entered and remained in the stepwise procedure. In discriminating European and North American salmon, the misclassification rate was only 0.35% with an overall bias of 0.2% in favour of European salmon. The discriminant analysis of the five European rivers resulted in a misclassification rate of 3.84%; the six North American rivers sampled

in 1968 resulted in a 5.52% misclassification rate; and finally the eight North American rivers sampled in 1969 gave a misclassification rate of 17.13%.

To verify these classification procedures, the jackknife classification technique (Section 3.5, Chapter 3) was used to determine the bias which may have resulted in basing classification decisions on that data set used to determine the classification functions. For the analysis of North American versus European data, the jackknife technique correctly classified 99.68% of the cases. This percentage ws actually 0.3% higher than the 99.65% originally classified correctly. For the analysis of the five European rivers, 95.33% were correctly classified using the jackknife technique. This was only 0.83% lower than the original classification rate of 96.16%. Thus, the jackknife procedure supports the original classification procedure.

Next, canonical variables were used as a means of classification instead of the original discriminating variables (Section 3.6, Chapter 3). Thus, by plotting the first two canonical variables with the plot of cases, it can graphically be seen how well the discriminating variables are classifying the cases. For all cases tested, results closely matched the previous misclassification rates.

Finally, a discrimination analysis based on the clustering principle was examined. In particular, the partitioning technique using the k sample means as initial cluster centres were used.

The results matched well with the discriminant analysis. However, the clustering procedure generally failed to separate approximately 10% more salmon than the discriminate analysis. This is because the variations among the variables are not considered in clustering techniques.

APPENDIX A

Table A1.1 Basic Statistics for European Data

	n	PREDOR	DORS	HEAD	POSTOR	LFTPECT	STNDLEN
Logan R.	117	59.731[1]	30.625	32.922	16.961	23.635	131.564
		4.181[2]	3.122	2.126	1.467	1.622	9.987
R. Almond	158	45.564	22.391	27.802	14.052	22.775	102.247
		3.962	2.850	2.002	1.112	1.580	9.689
R. Bayne	50	48.730	25.688	28.166	14.980	21.436	110.480
		3.598	3.272	2.097	1.187	1.626	8.811
R. Lee	98	54.454	30.515	31.329	16.538	24.980	125.653
		2.938	2.174	1.833	0.996	1.599	6.354
R. Usk	72	58.971	33.025	33.525	17.442	25.840	139.444
		5.144	3.394	2.523	1.321	1.993	12.529
Europe	495	52.942	27.825	30.575	15.818	23.725	120.053
		7.196	5.065	3.179	1.830	2.129	17.162

1 - mean

2 - standard deviation

Table A1.2 Basic Statistics for North American Data (1968 and 1969)

	n	PREDOR	DORS	HEAD	POSTOR	LFTPECT	STNDLEN
1969							
Maine	142	63.110[1]	35.199	36.101	18.570	27.6579	144.359
		3.948[2]	3.347	1.870	1.096	1.669	9.270
Miramichi R.	151	52.582	29.852	31.713	16.473	24.940	122.205
		4.932	3.318	2.858	1.514	1.789	11.036
Saint John R.	73	63.904	36.937	36.636	19.940	27.677	146.247
		6.370	4.132	3.215	1.843	2.691	14.736
Koksoak	130	58.402	34.057	32.704	16.822	26.639	135.515
		10.397	7.114	5.393	2.907	3.632	23.932
Indian R.	125	57.398	32.350	33.665	17.330	24.686	132.576
		4.305	3.050	2.211	1.244	1.481	10.242
Salmon R.	41	56.529	31.595	33.622	17.080	25.807	130.683
		8.828	6.349	5.003	2.700	3.048	21.501
Harry's R.	89	50.693	27.374	30.094	15.542	23.720	116.337
		4.675	3.711	2.517	1.428	1.626	11.296
Sand Hill R.	148	61.501	35.648	36.336	18.924	27.643	143.635
		4.707	3.647	2.226	1.391	1.637	11.345
North America	915	58.306	33.125	33.954	17.620	26.216	134.749
		7.612	5.268	3.874	2.183	2.653	17.606

1 - mean

2 - standard deviation

Table A1.2 (cont'd)

	n	PREDOR	DORS	HEAD	POSTOR	LFTPECT	STNDLEN
1968							
Maine	81	65.020	36.207	34.326	15.985	25.972	143.284
		4.509	3.619	2.080	1.283	1.892	10.532
Miramichi	147	51.989	28.847	30.931	15.090	23.220	117.048
		3.338	2.332	1.602	0.949	1.318	7.496
Saint John	148	58.968	34.339	33.357	16.070	24.517	136.628
		5.545	3.856	2.713	1.603	1.926	13.656
Indian R.	50	58.820	31.608	32.360	16.602	24.036	129.940
		4.223	2.837	2.141	1.191	1.576	9.859
Salmon R.	147	60.702	33.502	35.891	16.805	26.647	138.891
		3.616	2.920	1.799	1.001	1.343	8.594
Salmonier R.	151	56.501	32.210	33.266	15.189	24.383	128.901
		4.223	3.144	2.319	1.177	1.730	10.241
North America	724	58.055	32.663	33.379	15.864	24.788	131.783
		5.775	3.897	2.693	1.385	2.025	13.447

Figure A2.1 Boxplots of 1969 European Data
(All measurements in millimetres)

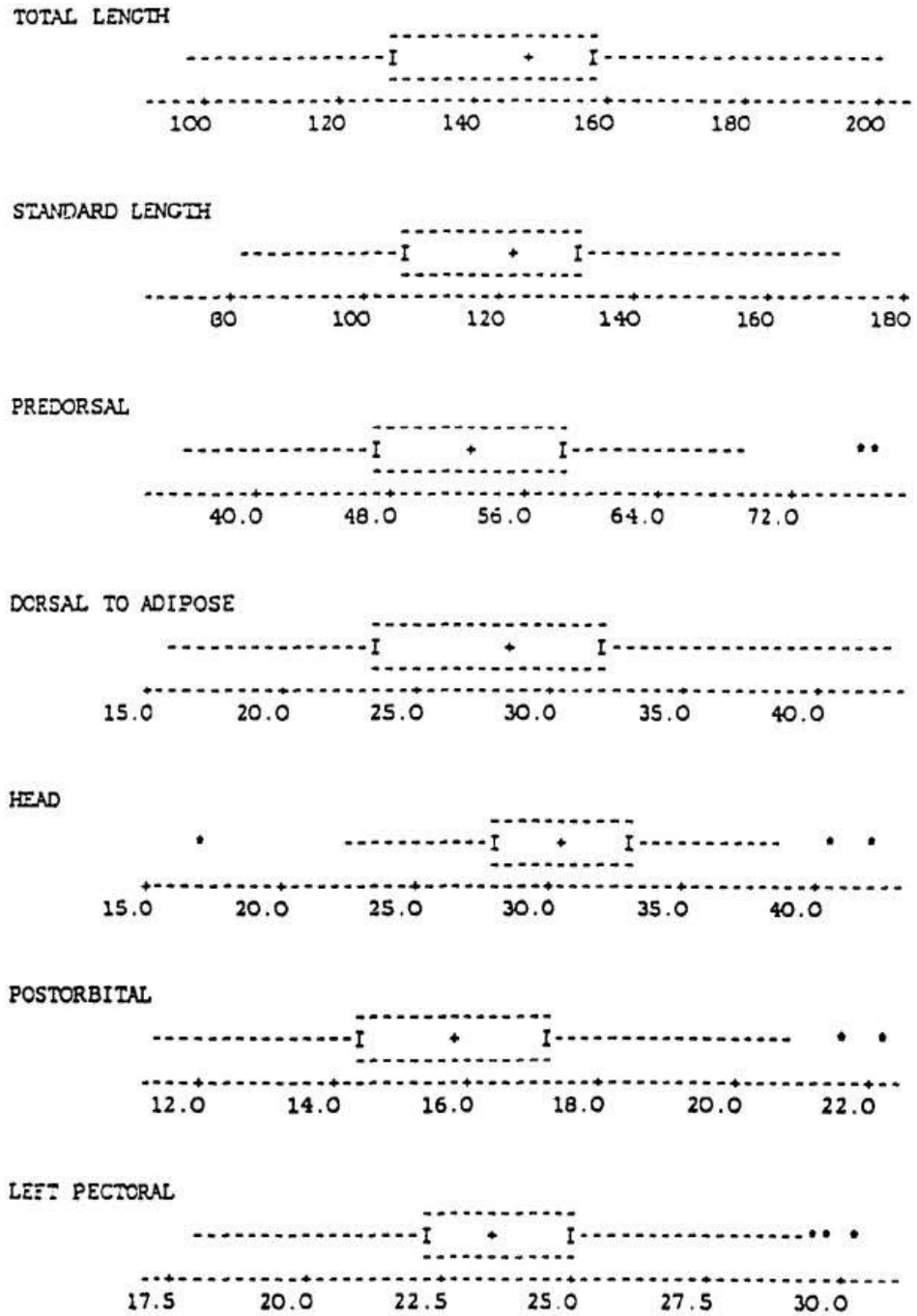


Figure A2.2 Boxplots of 1969 North American Data
(All measurements in millimetres)

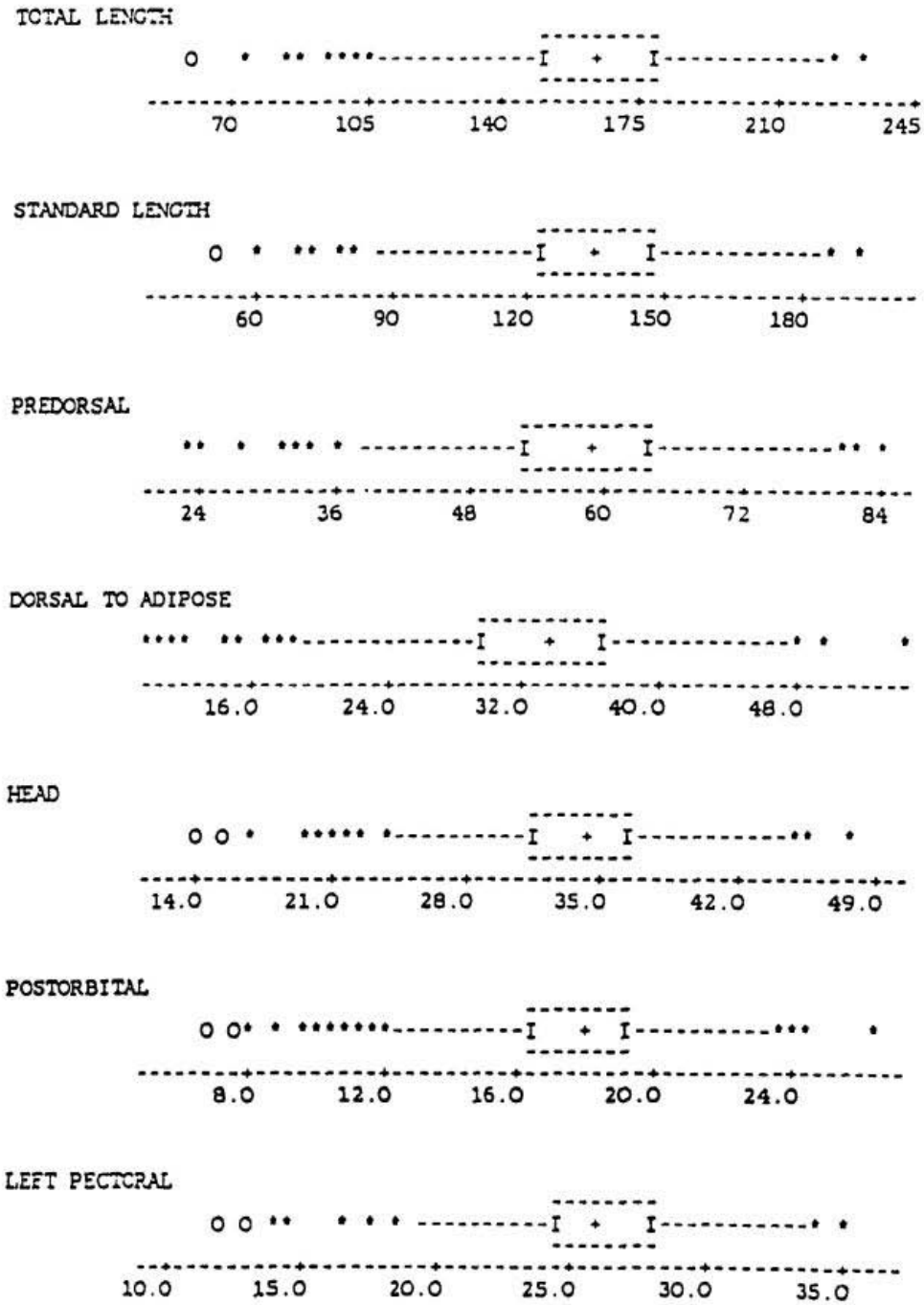


Figure A2.3 Boxplots of 1968 North American Data
(All measurements in millimetres)

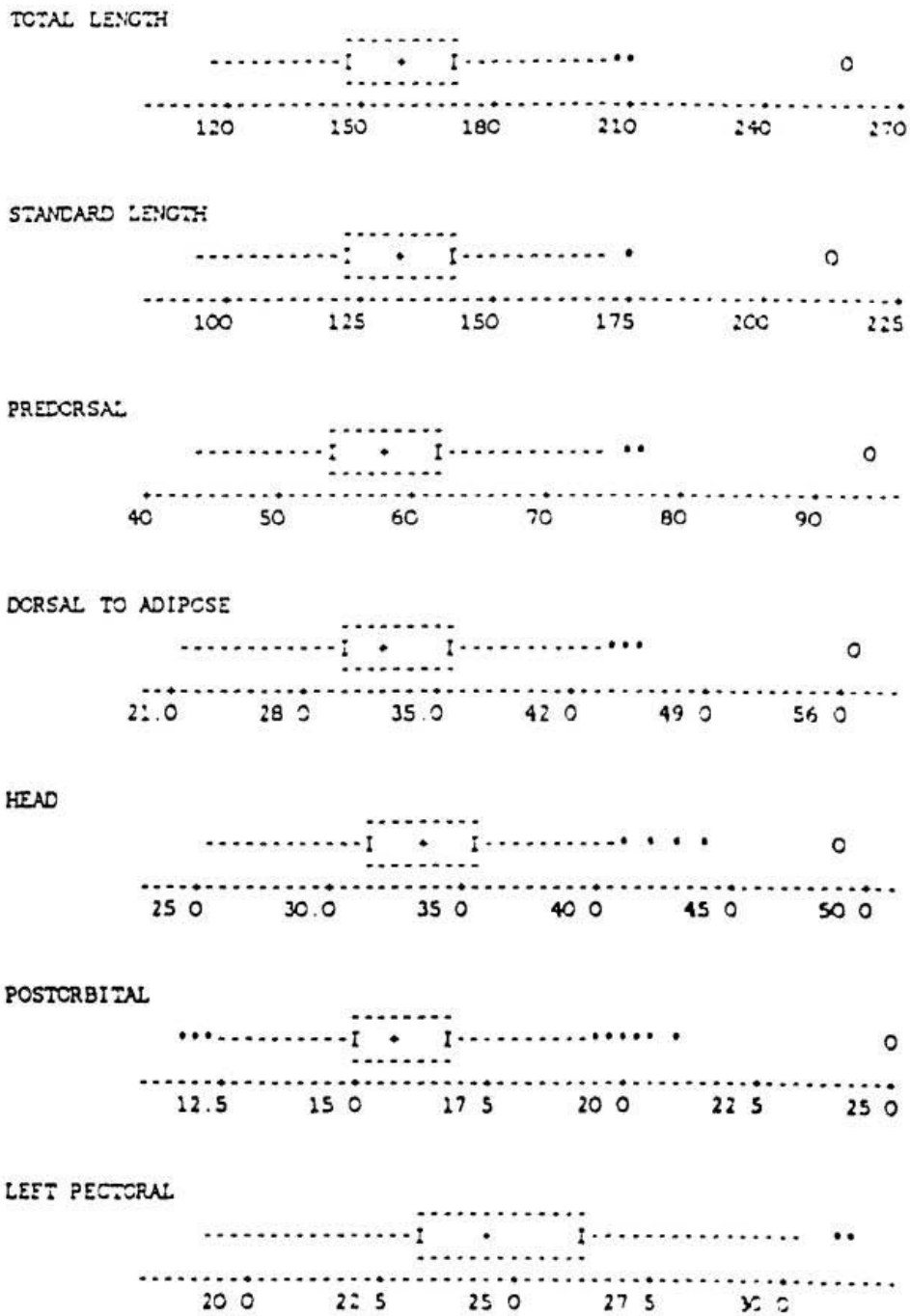
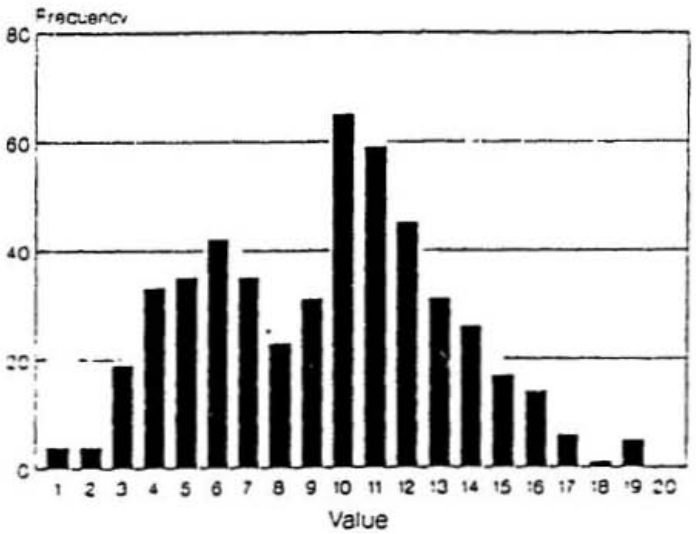


Figure A2.4 Character Distributions of 1969 European Specimens

a) Total Length



b) Standard Length

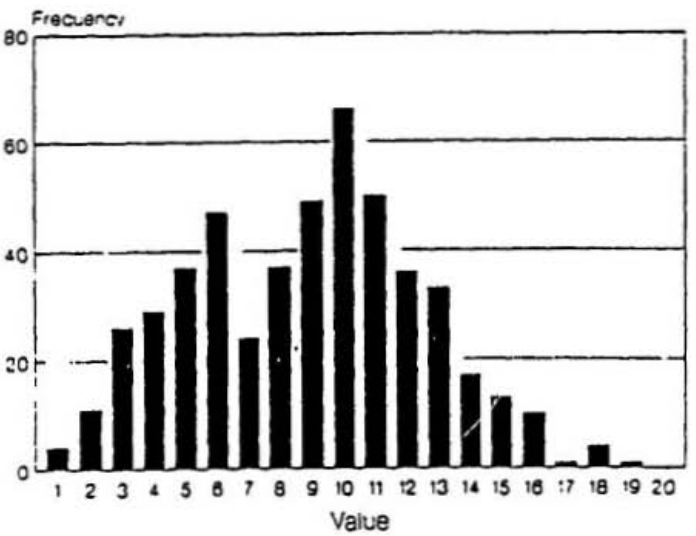
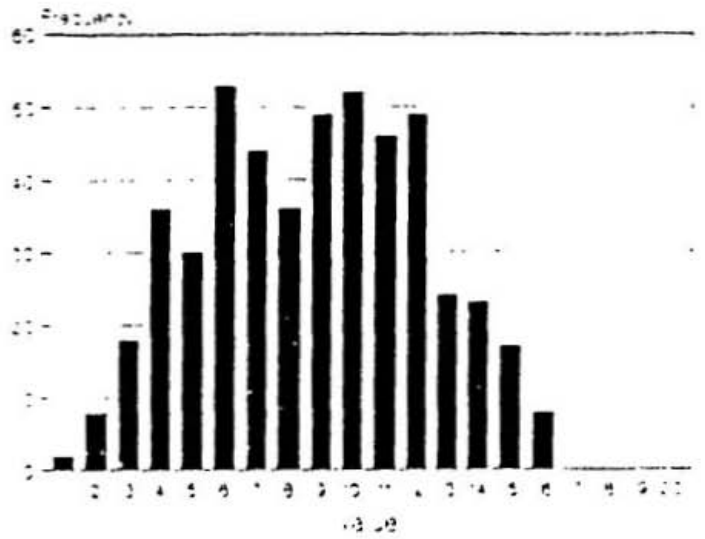


Figure A2.4 continued

c) Precorsal



d) Dorsal to Adipose

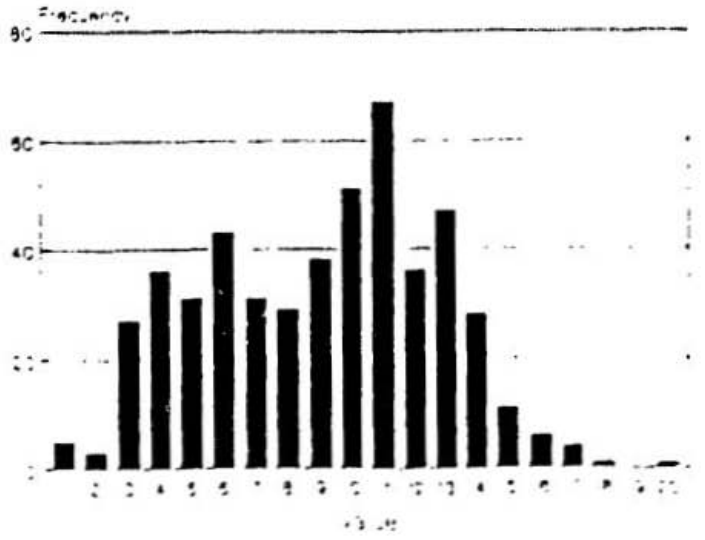
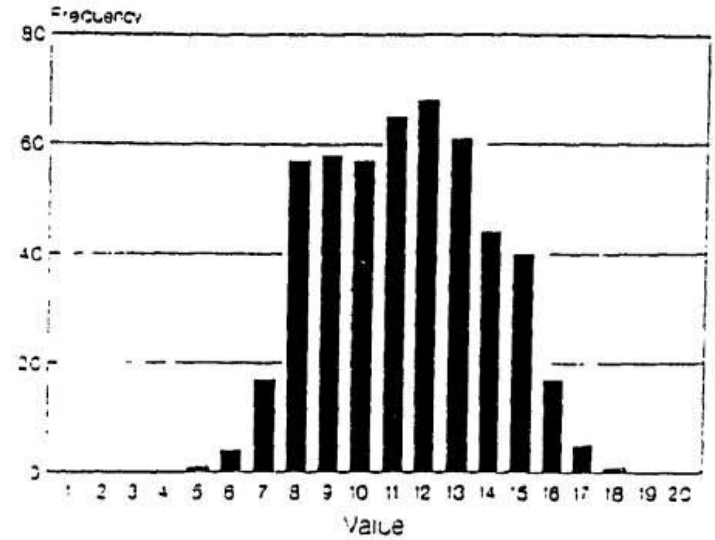


Figure A2.4 continued

e) Head



f) Postorbital

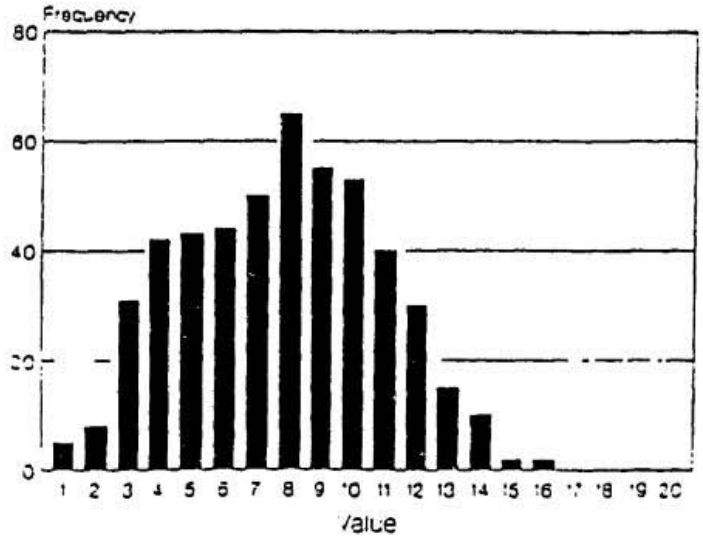


Figure A2.4 continued

g) Left Pectoral

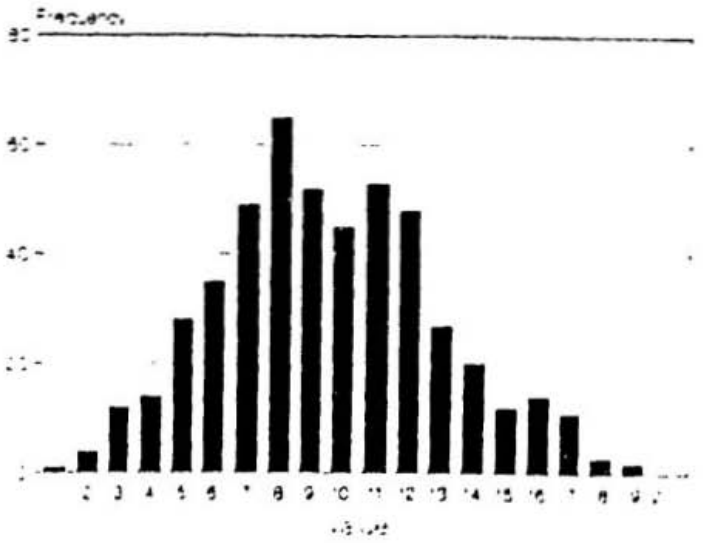
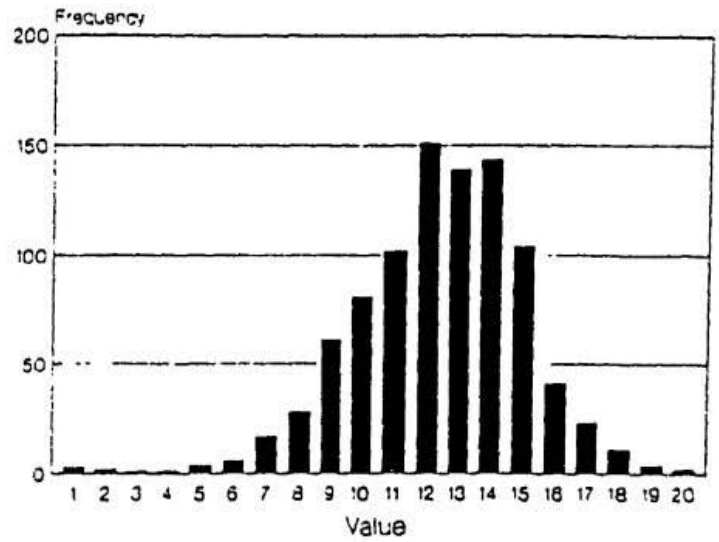


Figure A2.5 Character Distributions of 1969 North American Specimens

a) Total Length



b) Standard Length

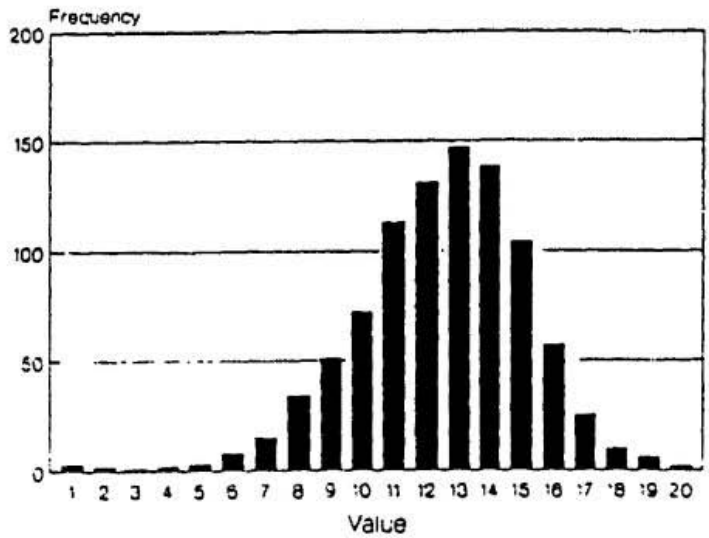
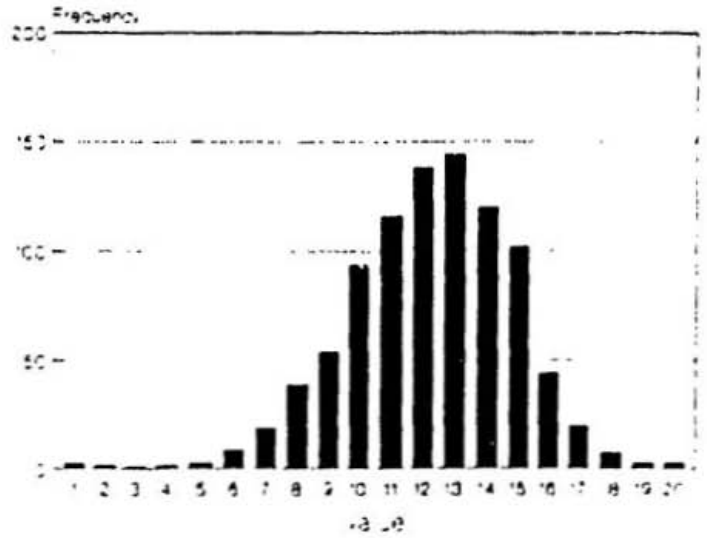


Figure A2.5 continued

c) Predorsal



d) Dorsal to Adipose

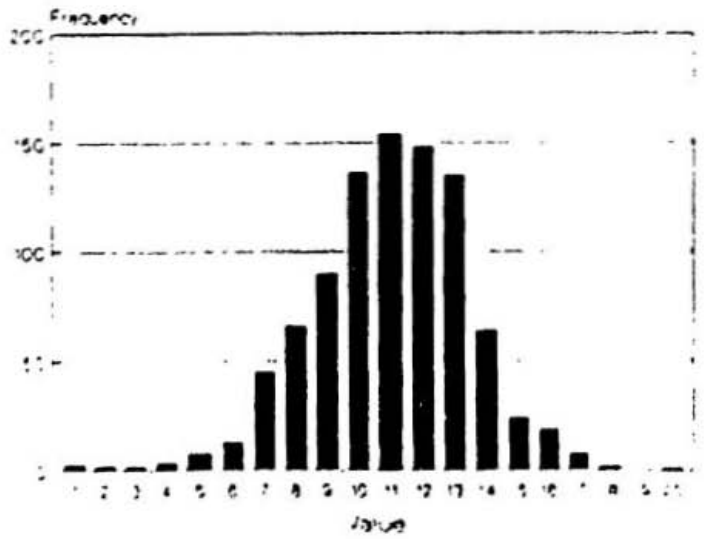
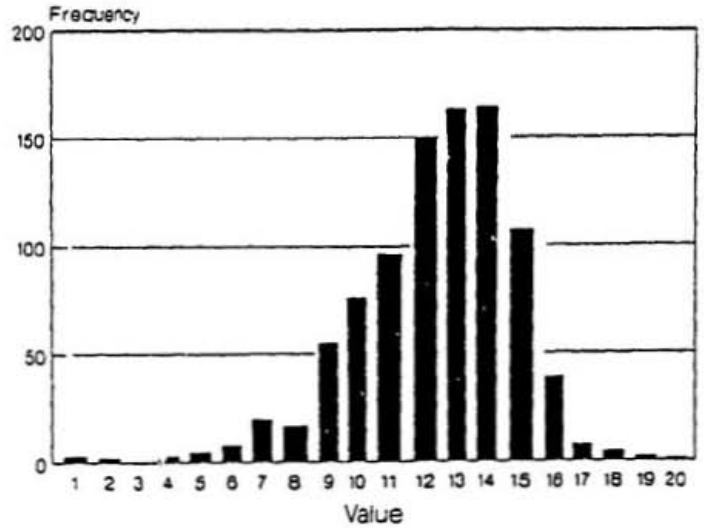


Figure A2.5 continued

e) Head



f) Postorbital

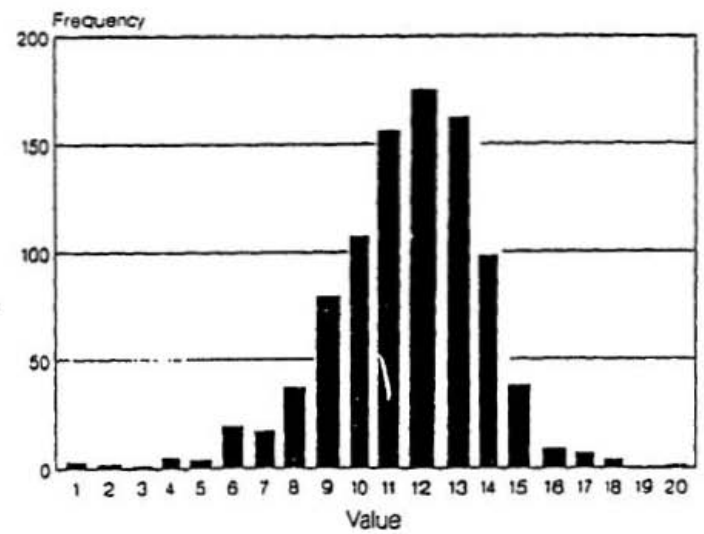


Figure A2.5 continued

g) Left Pectoral

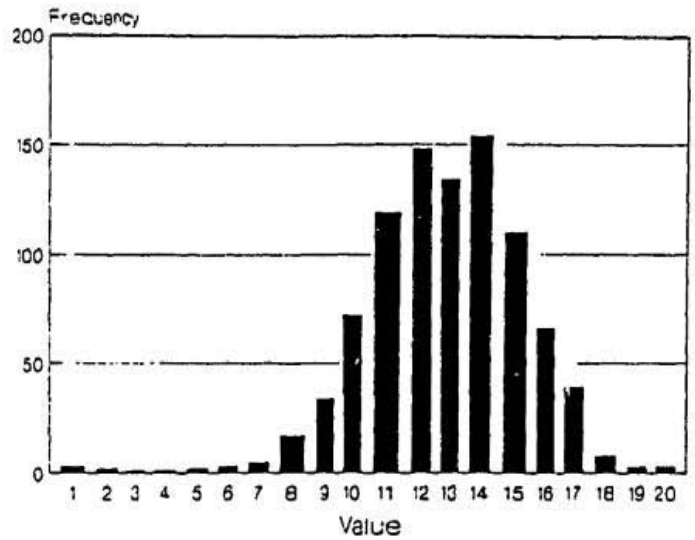
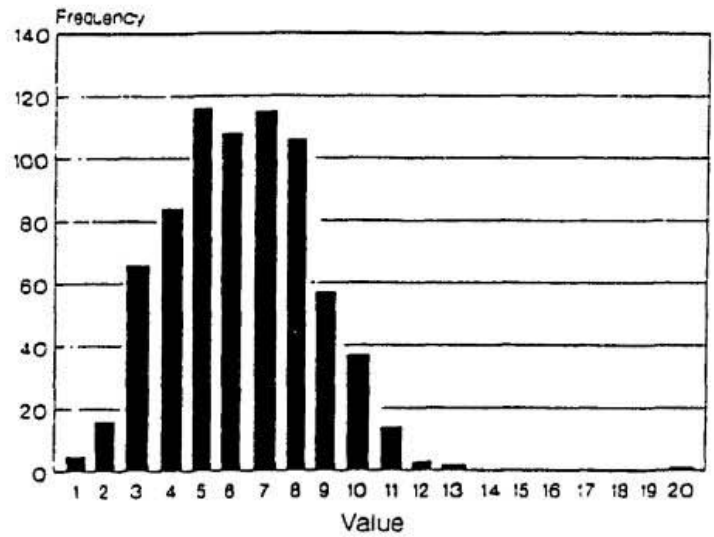


Figure A2.6 Character Distributions of 1968 North American Specimens

a) Total Length



b) Standard Length

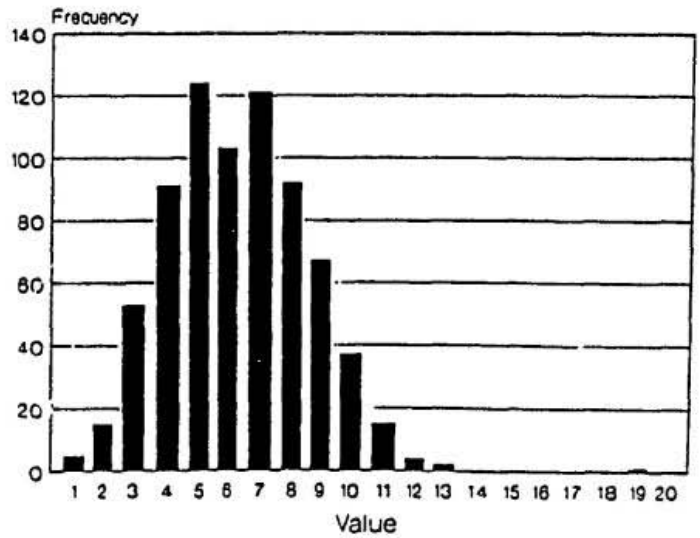
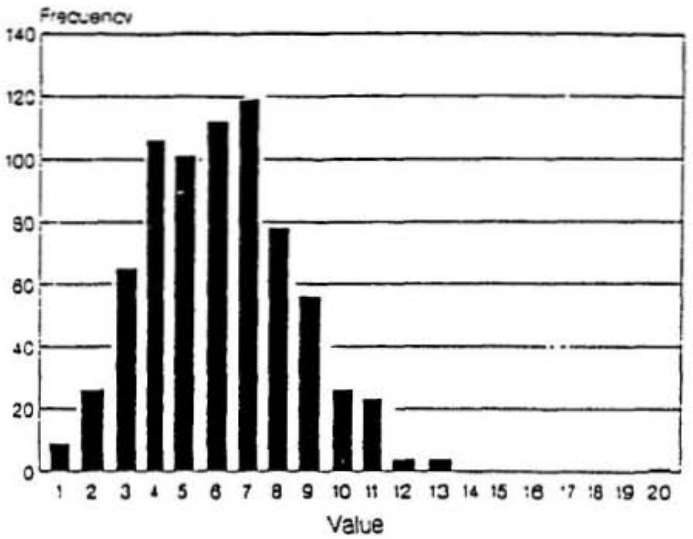


Figure A2.6 continued

c) Predorsal



d) Dorsal to Adipose

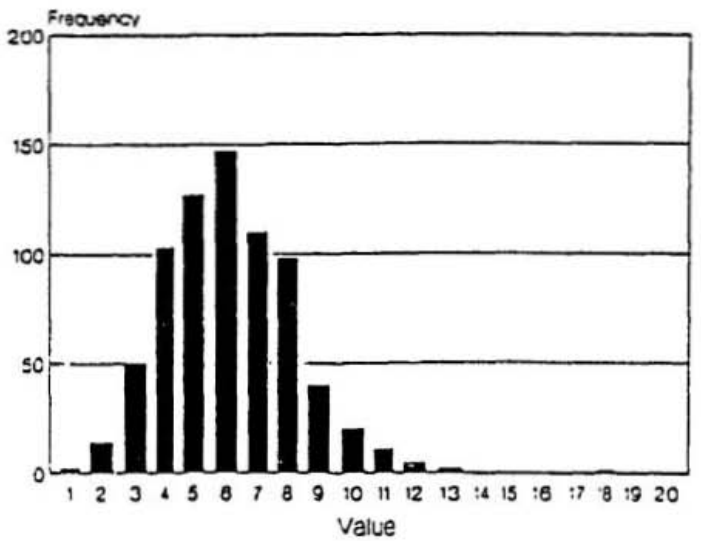
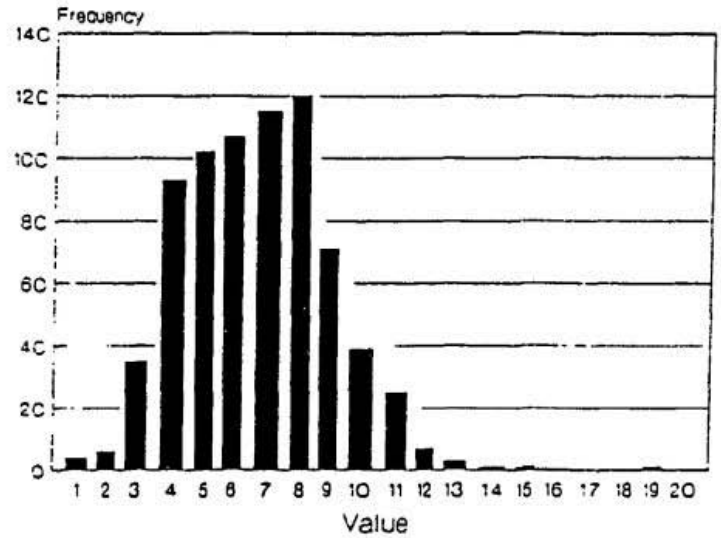


Figure A2.6 continued

e) Head



f) Postorbital

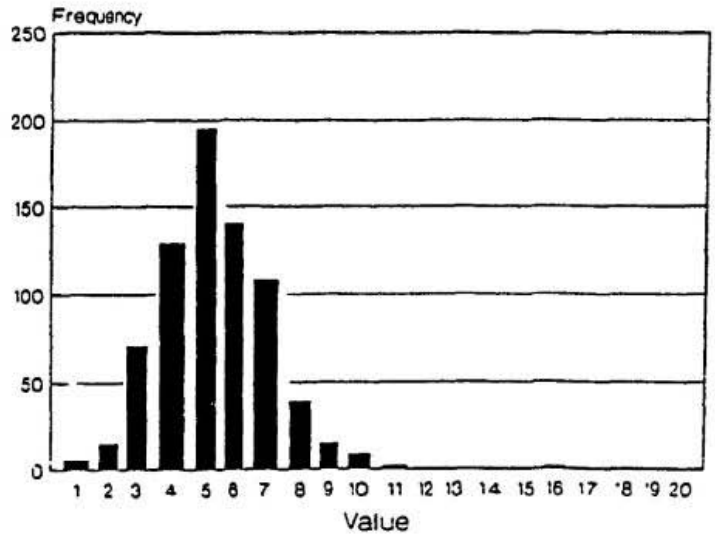


Figure A2.6 continued

g)

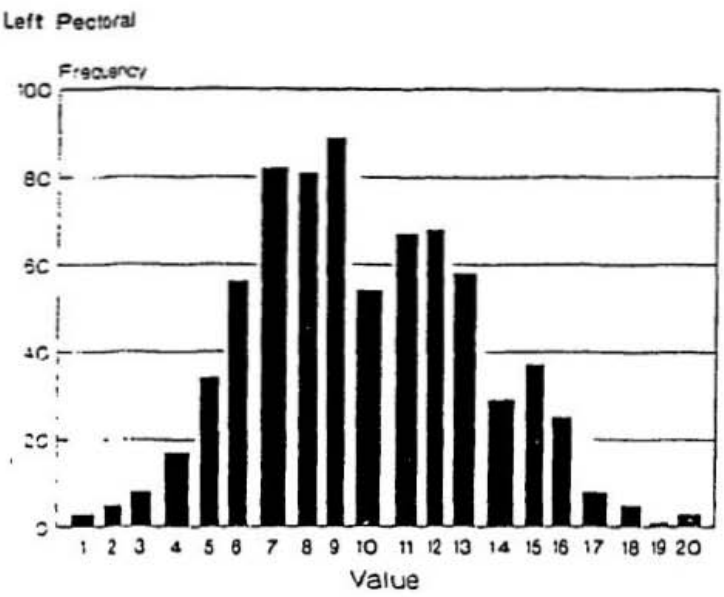


Table A2.1 Mid summeries, Spreads and Quotients for Europe (1969)

a) Total Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	128.0	158.0	143.00	30.00	1.11
E	119.0	167.0	143.00	48.00	1.04
D	114.0	176.0	145.00	62.00	1.01
C	111.0	180.5	145.75	69.50	0.93
B	108.5	190.5	149.5	82.00	0.95
A	103.0	195.5	149.25	92.50	0.95
Z	100.5	196.0	148.25	95.50	0.90
Y	98.5	197.5	148.0	99.00	0.86

b) Standard Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	106.0	132.0	119.00	26.00	1.09
E	98.0	140.0	119.00	42.00	1.03
D	94.0	147.0	120.50	53.00	0.98
C	91.0	152.5	121.75	61.50	0.94
B	88.5	160.5	124.50	72.00	0.95
A	85.0	164.5	124.75	79.50	0.93
Z	82.5	167.0	124.75	84.50	0.90
Y	81.5	169.0	125.25	87.50	0.86

c) Predorsal

Depth	Lower	Upper	Mid	Spread	Quotient
H	47.20	58.55	52.88	11.35	1.13
E	44.00	61.50	52.75	17.50	1.02
D	42.10	64.30	53.20	22.20	0.96
C	41.10	66.40	53.75	25.30	0.91
B	39.35	67.65	53.50	28.30	0.88
A	38.70	68.55	53.63	29.85	0.82
Z	37.40	72.60	55.00	35.20	0.88
Y	36.30	76.60	56.45	40.30	0.93

Table A2.1 (cont'd)

d) Dorsal to Adipose

Depth	Lower	Upper	Mid	Spread	Quotient
H	23.50	31.85	27.68	8.35	1.19
E	21.20	33.70	27.45	12.50	1.05
D	20.00	35.00	27.50	15.00	0.94
C	19.35	36.25	27.80	16.90	0.91
B	18.35	38.60	28.48	20.25	0.91
A	17.00	39.70	28.35	22.70	0.90
Z	16.35	41.40	28.88	25.05	0.91
Y	16.05	42.25	29.15	26.20	0.88

e) Head

Depth	Lower	Upper	Mid	Spread	Quotient
H	28.00	32.90	30.45	4.90	1.07
E	26.50	34.70	30.60	8.20	1.06
D	26.00	35.70	30.85	9.70	0.94
C	25.45	36.35	30.90	10.90	0.87
B	24.70	37.45	31.08	12.75	0.88
A	23.65	38.05	30.85	14.40	0.88
Z	23.00	39.50	31.25	16.50	0.92
Y	19.75	41.20	30.48	21.45	1.11

f) Postorbital

Depth	Lower	Upper	Mid	Spread	Quotient
H	14.40	17.20	15.80	2.80	1.07
E	13.50	18.00	15.75	4.50	1.01
D	13.10	18.70	15.90	5.60	0.94
C	12.85	19.30	16.08	6.45	0.90
B	12.40	19.50	15.95	7.10	0.85
A	12.00	20.40	16.20	8.40	0.90
Z	11.55	21.15	16.35	9.60	0.93
Y	11.45	21.85	16.65	10.40	0.94

Table A2.1 (cont'd)

g) *Left Pectoral*

Depth	Lower	Upper	Mid	Spread	Quotient
H	22.20	25.10	23.65	2.90	0.98
E	21.30	26.30	23.80	5.00	0.99
D	20.70	27.50	24.10	6.80	1.01
C	19.90	28.30	24.10	8.40	1.03
B	19.50	28.75	24.13	9.25	0.98
A	19.30	29.55	24.43	10.25	0.96
Z	19.00	30.00	24.50	11.00	0.94
Y	18.55	30.25	24.40	11.70	0.93

Table A2.2 Mid summaries, Spreads and Quotients for North America (1969)

a) Total Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	149.0	177.5	163.25	28.50	0.93
E	136.0	186.0	161.00	50.00	0.96
D	128.0	193.0	160.50	65.00	0.93
C	118.0	201.0	159.50	83.00	0.98
B	110.0	207.0	158.50	97.00	0.99
A	95.0	212.0	153.50	117.00	1.06
Z	72.5	219.0	145.75	146.50	1.21
Y	61.0	223.5	142.25	162.50	1.24
X	60.5	228.0	144.25	167.50	1.19

Table A2.2 (cont'd)

b) Standard Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	123.0	146.0	135.00	24.00	0.94
E	112.0	154.0	133.00	42.00	0.96
D	106.0	161.0	133.50	55.00	0.94
C	97.0	166.5	131.75	69.50	0.98
B	90.0	171.0	130.50	81.00	0.99
A	78.0	177.0	127.50	99.00	1.08
Z	59.5	181.5	120.50	122.00	1.21
Y	50.0	184.0	117.00	134.00	1.23
X	50.0	188.0	119.00	138.00	1.18

c) Predorsal

Depth	Lower	Upper	Mid	Spread	Quotient
H	53.10	64.00	58.55	10.90	0.99
E	48.70	66.70	57.70	18.00	0.96
D	45.50	69.05	57.28	23.55	0.94
C	41.85	71.70	56.78	29.85	0.98
B	39.80	73.90	56.85	34.10	0.97
A	33.60	76.00	54.80	42.40	1.07
Z	27.45	78.40	52.93	50.95	1.17
Y	24.00	81.55	52.78	57.55	1.23
X	23.20	82.85	53.03	59.65	1.18

d) Dorsal to Adipose

Depth	Lower	Upper	Mid	Spread	Quotient
H	29.70	36.70	33.20	7.00	0.93
E	26.90	38.70	32.80	11.80	0.92
D	24.50	40.65	32.58	16.15	0.94
C	22.60	43.45	33.03	20.85	1.00
B	19.80	45.00	32.40	25.20	1.04
A	16.90	46.00	31.45	29.10	1.07
Z	12.35	46.95	29.65	34.60	1.16
Y	10.60	48.85	29.73	38.25	1.19
X	9.95	52.00	30.98	42.05	1.22

Table A2.2 (cont'd)

e) Head

Depth	Lower	Upper	Mid	Spread	Quotient
H	31.55	36.60	34.08	5.05	0.89
E	29.10	38.00	33.55	8.90	0.92
D	27.45	39.00	33.23	11.55	0.90
C	24.75	40.10	32.43	15.35	0.98
B	22.70	41.40	32.05	18.70	1.03
A	20.10	42.50	31.30	22.40	1.10
Z	16.90	43.85	30.38	26.95	1.21
Y	14.60	44.90	29.75	30.30	1.26
X	14.05	46.40	30.23	32.35	1.25

f) Postorbital

Depth	Lower	Upper	Mid	Spread	Quotient
H	16.30	19.10	17.70	2.80	0.88
E	15.10	20.00	17.55	4.90	0.90
D	13.90	20.60	17.25	6.70	0.93
C	12.50	21.40	16.95	8.90	1.02
B	11.70	22.00	16.85	10.30	1.01
A	10.00	23.00	16.50	13.00	1.14
Z	8.30	23.60	15.95	15.30	1.22
Y	7.20	24.20	15.70	17.00	1.25
X	6.85	25.35	16.10	18.50	1.27

g) Left Pectoral

Depth	Lower	Upper	Mid	Spread	Quotient
H	24.40	28.00	26.20	3.60	0.93
E	23.00	29.30	26.15	6.30	0.95
D	22.00	30.20	26.10	8.20	0.93
C	21.05	31.00	26.03	9.95	0.93
B	19.60	31.40	25.50	11.80	0.95
A	16.60	32.30	24.45	15.70	1.13
Z	14.00	33.25	23.63	19.25	1.26
Y	12.45	34.55	23.50	22.10	1.34
X	12.00	34.95	23.48	22.95	1.29

Table A2.3 Mid summaries, Spreads and Quotients for North America (1968)

a) Total Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	147.0	170.0	158.50	23.00	0.71
E	140.0	179.0	159.50	39.00	0.71
D	135.0	185.0	160.00	50.00	0.68
C	132.0	190.5	161.25	58.50	0.66
B	128.0	195.0	161.50	67.00	0.65
A	125.5	199.5	162.50	74.00	0.64
Z	122.0	206.0	164.00	84.00	0.66
Y	122.0	209.0	165.50	87.00	0.63

b) Standard Length

Depth	Lower	Upper	Mid	Spread	Quotient
H	122.0	142.0	132.00	20.00	0.74
E	115.0	148.0	131.50	33.00	0.72
D	112.0	153.0	132.50	41.00	0.67
C	109.0	157.5	133.25	48.50	0.65
B	106.0	164.0	135.00	58.00	0.68
A	103.0	166.5	134.75	63.50	0.66
Z	100.0	169.5	134.75	69.50	0.65
Y	100.0	174.0	137.00	74.00	0.64

c) Predorsal

Depth	Lower	Upper	Mid	Spread	Quotient
H	53.80	62.00	57.90	8.20	0.70
E	51.30	65.20	58.25	13.90	0.69
D	49.65	67.90	58.78	18.25	0.68
C	48.55	70.00	59.28	21.45	0.66
B	46.90	71.10	59.00	24.20	0.64
A	46.45	73.40	59.93	26.95	0.64
Z	45.15	76.15	60.65	31.00	0.67
Y	45.00	76.80	60.90	31.80	0.63

Table A2.3 (cont'd)

d) Dorsal to Adipose

Depth	Lower	Upper	Mid	Spread	Quotient
H	29.80	35.50	32.65	5.70	0.71
E	28.30	37.30	32.80	9.00	0.66
D	27.05	39.05	33.05	12.00	0.66
C	26.10	40.70	33.40	14.60	0.66
B	25.50	42.30	33.90	16.80	0.66
A	24.50	44.00	34.25	19.50	0.68
Z	23.80	45.35	34.58	21.55	0.68
Y	23.20	45.70	34.45	22.50	0.66

e) Head

Depth	Lower	Upper	Mid	Spread	Quotient
H	31.40	35.30	33.35	3.90	0.76
E	30.20	36.70	33.45	6.50	0.74
D	29.50	37.80	33.65	8.30	0.71
C	28.90	38.85	33.88	9.95	0.70
B	28.40	39.70	34.05	11.30	0.69
A	27.45	40.75	34.10	13.30	0.73
Z	26.70	42.35	34.53	15.65	0.77
Y	25.90	43.90	34.90	18.00	0.82

f) Postorbital

Depth	Lower	Upper	Mid	Spread	Quotient
H	15.00	16.80	15.90	1.80	0.66
E	14.20	17.50	15.85	3.30	0.71
D	13.85	18.05	15.95	4.20	0.67
C	13.50	18.80	16.15	5.30	0.70
B	13.20	19.50	16.35	6.30	0.72
A	12.60	19.95	16.28	7.35	0.75
Z	12.20	20.40	16.30	8.20	0.76
Y	11.80	21.00	16.40	9.20	0.79

Table A2.3 (cont'd)

g) Left Pectoral

Depth	Lower	Upper	Mid	Spread	Quotient
H	23.30	26.30	24.80	3.00	1.07
E	22.50	27.30	24.90	4.80	1.00
D	22.00	28.20	25.10	6.20	0.97
C	21.30	28.70	25.00	7.40	0.96
B	20.70	29.20	24.95	8.50	0.95
A	20.30	29.75	25.03	9.45	0.94
Z	20.00	30.65	25.33	10.65	0.97
Y	19.70	31.00	25.35	11.30	0.95

Table A3.1

D^2, W and p-values

D^2 - First entry
W - Second entry
p-value - Third entry

	Europe (1969)
North America (1969)	10,598.6[1] 2,117.7[2] 0.0000[3]

Table A3.2

	Logan R.	R. Almond	R. Boyne	R. Lee
	11,415.3			
R. Almond	2249.6			
	0.0000			
	3811.60	496.190		
R. Bayne	743.84	97.311		
	0.0000	0.0000		
	1453.19	4988.51	1422.52	
R. Lee	285.18	981.99	276.71	
	0.0000	0.0000	0.0000	
	371.169	8856.38	3513.16	852.959
R. Usk	72.646	1740.2	679.21	166.53
	0.0000	0.0000	0.0000	0.0000

Table A3.3

	Maine	Miramichi	Saint John	Indian R.	Salmon R.
	8876.61				
Miramichi	1743.9				
	0.0000				
	1605.34	4521.63			
Saint John	315.41	891.98			
	0.0000	0.0000			
	2055.25	1568.52	458.763*		
Indian R.	369.65	307.27	89.880		
	0.0000	0.0000	0.0000		
	1398.45	7175.80	963.858	1405.79	
Salmon R.	274.74	1415.5	190.14	275.39	
	0.0000	0.0000	0.0000	0.0000	
	3422.83	2487.92	898.552*	935.249*	1673.16
Salmonier R.	672.66	490.86	177.29	183.29	330.11
	0.0000	0.0000	0.0000	0.0000	0.0000

Table A3.4

	Maine	Miramichi	Saint John	Koksoak R.	Indian R.	Salmon R.	Harry's R
	22,208.8						
Miramichi	4380.7						
	0.0000						
	83.758*	17,348.6					
Saint John	14.437	3407.2					
	0.0000	0.0000					
	4124.20	6481.42	3500.57				
Koksoak R.	812.62	1277.7	764.59				
	0.0000	0.0000	0.0000				
	5947.78	4346.20	5378.77	176.175			
Indian R.	1171.6	856.55	1053.8	34.678			
	0.0000	0.0000	0.0000	0.0000			
	3804.37	1385.93	3999.07	302.152	64.432*		
Salmon R.	744.06	271.35	771.25	59.000	12.572		
	0.0000	0.0000	0.0000	0.0000	0.0004		
	23,182.94	548.671	19,387.2	8636.19	6431.35	2665.03	
Harry's R.	4555.6	107.89	3780.5	1695.4	1262.0	516.35	
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	513.887	16,274.2	776.941	1821.60	3126.90	2189.59	17,833.6
Sand Hill R.	101.35*	3211.0	152.55*	359.04	616.15	428.55	3506.0
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

APPENDIX B

Jackknife of Europe '69 vs N. America '69

All variables entered with covariate STNDLEN

n = 1410
500 samples
p = 0.9858

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
8	15	0	100.00
7	16	0	100.00
5	12	0	100.00
5	21	0	100.00
9	11	0	100.00
10	12	0	100.00
12	12	0	100.00
5	11	1	93.75
8	12	0	100.00
9	12	1	95.24
6	13	0	100.00
8	12	0	100.00
9	11	0	100.00
4	9	0	100.00
11	12	0	100.00
6	14	0	100.00
5	13	1	94.44
5	7	0	100.00
4	12	0	100.00
7	15	0	100.00
6	16	0	100.00
8	10	0	100.00
6	19	0	100.00
8	13	0	100.00
3	13	0	100.00
3	8	0	100.00
6	16	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
4	12	0	100.00
5	10	0	100.00
8	8	0	100.00
5	7	0	100.00
9	11	0	100.00
5	19	0	100.00
5	12	0	100.00
7	14	0	100.00
7	18	0	100.00
8	12	0	100.00
8	13	0	100.00
4	6	0	100.00
7	10	0	100.00
5	20	0	100.00
4	9	0	100.00
8	11	0	100.00
11	13	0	100.00
5	11	0	100.00
7	11	0	100.00
8	13	0	100.00
3	18	0	100.00
4	6	1	90.00
15	10	0	100.00
8	14	0	100.00
9	21	0	100.00
5	12	1	94.12
6	19	0	100.00
5	3	0	100.00
10	10	0	100.00
6	20	0	100.00
12	15	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
10	9	0	100.00
6	9	0	100.00
10	6	0	100.00
8	14	0	100.00
8	25	0	100.00
10	14	0	100.00
10	10	0	100.00
8	8	0	100.00
5	17	1	95.45
7	11	0	100.00
2	11	0	100.00
6	10	0	100.00
10	10	0	100.00
9	18	0	100.00
7	13	0	100.00
11	19	0	100.00
3	18	0	100.00
7	15	0	100.00
5	9	0	100.00
12	11	0	100.00
5	11	0	100.00
8	19	0	100.00
8	13	0	100.00
7	16	0	100.00
5	8	0	100.00
5	13	0	100.00
9	19	1	96.43
5	9	0	100.00
8	17	0	100.00
4	8	0	100.00
8	10	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
6	15	0	100.00
8	8	0	100.00
11	14	0	100.00
10	16	0	100.00
10	6	0	100.00
7	19	0	100.00
6	12	0	100.00
8	19	1	96.30
4	21	0	100.00
10	15	0	100.00
8	12	0	100.00
12	15	0	100.00
6	11	0	100.00
10	15	0	100.00
13	19	0	100.00
4	19	0	100.00
8	16	0	100.00
9	11	0	100.00
8	17	0	100.00
7	9	0	100.00
12	12	0	100.00
5	16	0	100.00
6	17	0	100.00
7	8	0	100.00
5	13	0	100.00
6	6	0	100.00
4	18	0	100.00
7	18	0	100.00
4	12	0	100.00
7	20	0	100.00
7	11	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
12	12	0	100.00
9	16	0	100.00
6	17	0	100.00
8	20	0	100.00
10	6	0	100.00
6	9	0	100.00
8	15	0	100.00
5	12	0	100.00
6	12	0	100.00
9	15	0	100.00
11	14	0	100.00
4	12	0	100.00
7	17	0	100.00
8	20	0	100.00
6	13	0	100.00
7	15	0	100.00
8	16	0	100.00
8	19	0	100.00
4	6	0	100.00
11	10	0	100.00
5	12	0	100.00
9	11	0	100.00
8	10	0	100.00
8	13	0	100.00
3	10	0	100.00
6	7	0	100.00
7	12	0	100.00
7	8	0	100.00
5	13	0	100.00
6	17	0	100.00
4	8	1	91.67

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
9	13	0	100.00
11	6	0	100.00
5	12	0	100.00
10	10	0	100.00
8	8	0	100.00
7	18	0	100.00
13	8	0	100.00
6	6	0	100.00
6	25	0	100.00
7	13	0	100.00
4	8	0	100.00
5	17	0	100.00
10	15	0	100.00
11	13	0	100.00
9	23	0	100.00
7	16	0	100.00
7	14	0	100.00
8	12	0	100.00
7	6	0	100.00
10	12	0	100.00
5	13	1	94.44
10	9	0	100.00
7	15	0	100.00
10	16	0	100.00
5	10	0	100.00
6	6	0	100.00
10	15	1	96.00
5	12	0	100.00
3	20	0	100.00
8	17	0	100.00
6	16	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
9	12	0	100.00
6	17	0	100.00
8	8	0	100.00
5	7	0	100.00
3	13	0	100.00
5	9	0	100.00
9	6	0	100.00
7	13	0	100.00
7	6	0	100.00
7	15	0	100.00
3	12	0	100.00
5	11	0	100.00
8	10	0	100.00
12	12	0	100.00
6	15	0	100.00
8	9	0	100.00
11	16	0	100.00
6	15	0	100.00
6	15	0	100.00
4	20	0	100.00
5	14	0	100.00
7	12	0	100.00
8	10	0	100.00
9	15	0	100.00
9	19	0	100.00
5	7	0	100.00
10	20	0	100.00
10	16	1	96.15
10	12	0	100.00
6	14	1	95.00
5	9	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
10	14	1	100.00
10	9	0	100.00
8	16	0	100.00
11	10	0	100.00
9	11	0	100.00
5	15	0	100.00
7	15	0	100.00
6	8	0	100.00
2	18	0	100.00
11	12	0	100.00
6	14	0	100.00
5	7	0	100.00
8	14	0	100.00
4	11	0	100.00
9	12	0	100.00
3	9	0	100.00
9	10	0	100.00
2	16	0	100.00
9	17	0	100.00
9	13	0	100.00
7	9	0	100.00
7	11	0	100.00
6	16	0	100.00
4	14	0	100.00
5	20	0	100.00
10	6	0	100.00
9	11	0	100.00
10	13	0	100.00
8	12	0	100.00
8	9	0	100.00
16	10	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
11	9	0	100.00
4	15	0	100.00
6	18	0	100.00
8	8	0	100.00
5	23	0	100.00
2	15	0	100.00
4	5	0	100.00
7	10	0	100.00
11	14	1	96.00
5	15	0	100.00
5	9	0	100.00
8	15	1	95.65
8	7	0	100.00
6	12	0	100.00
11	8	0	100.00
6	16	0	100.00
4	14	0	100.00
6	12	0	100.00
7	10	0	100.00
8	14	0	100.00
3	21	0	100.00
5	14	0	100.00
8	15	0	100.00
10	17	0	100.00
10	10	1	95.00
4	14	1	94.44
7	16	0	100.00
9	10	0	100.00
9	8	0	100.00
6	10	0	100.00
7	9	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
5	20	0	100.00
11	12	0	100.00
6	17	0	100.00
12	10	0	100.00
4	12	0	100.00
6	19	0	100.00
6	14	0	100.00
6	8	0	100.00
6	11	0	100.00
8	14	0	100.00
10	10	0	100.00
6	10	0	100.00
11	10	0	100.00
6	15	0	100.00
5	13	0	100.00
5	14	0	100.00
9	15	0	100.00
4	5	0	100.00
6	10	0	100.00
6	11	0	100.00
7	15	0	100.00
6	9	0	100.00
8	9	1	94.12
5	14	0	100.00
7	18	0	100.00
9	19	0	100.00
11	16	0	100.00
12	11	1	95.65
6	11	0	100.00
8	14	0	100.00
8	24	1	96.88

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
11	11	0	100.00
7	10	0	100.00
6	8	0	100.00
4	14	1	94.44
3	13	0	100.00
5	9	0	100.00
9	17	0	100.00
7	9	0	100.00
6	9	0	100.00
6	14	0	100.00
6	17	0	100.00
12	11	0	100.00
7	13	0	100.00
1	14	0	100.00
2	14	0	100.00
8	13	0	100.00
11	15	0	100.00
8	18	0	100.00
8	18	1	96.15
5	17	0	100.00
4	10	0	100.00
7	16	0	100.00
6	12	0	100.00
7	7	0	100.00
5	9	0	100.00
6	16	0	100.00
7	7	0	100.00
7	17	0	100.00
4	17	0	100.00
7	10	0	100.00
9	21	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
7	8	0	100.00
9	10	0	100.00
6	15	0	100.00
6	13	0	100.00
10	16	0	100.00
3	12	0	100.00
7	11	0	100.00
4	20	0	100.00
2	10	0	100.00
10	6	0	100.00
10	16	0	100.00
7	18	0	100.00
13	12	0	100.00
8	13	0	100.00
11	12	0	100.00
9	16	0	100.00
5	9	0	100.00
8	15	0	100.00
2	11	0	100.00
7	16	0	100.00
7	9	0	100.00
5	11	0	100.00
8	14	0	100.00
2	15	0	100.00
5	10	0	100.00
6	10	0	100.00
11	10	1	95.24
9	13	1	95.45
6	18	0	100.00
7	12	0	100.00
11	13	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
9	16	0	100.00
5	17	0	100.00
4	8	0	100.00
8	17	0	100.00
6	15	0	100.00
3	12	0	100.00
9	12	0	100.00
6	9	0	100.00
6	14	0	100.00
7	15	0	100.00
9	14	0	100.00
8	13	1	95.45
5	14	0	100.00
4	13	1	95.24
12	20	0	100.00
7	11	0	100.00
3	11	0	100.00
6	15	0	100.00
5	12	0	100.00
9	11	0	100.00
9	20	0	100.00
6	13	0	100.00
6	12	0	100.00
12	15	0	100.00
5	11	0	100.00
5	15	0	100.00
5	18	0	100.00
5	8	0	100.00
8	15	0	100.00
2	14	0	100.00
9	13	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
5	14	0	100.00
8	11	0	100.00
7	12	0	100.00
9	20	0	100.00
11	10	1	95.24
9	6	0	100.00
4	10	0	100.00
2	20	0	100.00
4	9	0	100.00
9	17	0	100.00
13	18	0	100.00
5	12	0	100.00
12	16	0	100.00
7	13	0	100.00
4	14	0	100.00
8	6	0	100.00
6	12	0	100.00
8	13	0	100.00
11	16	0	100.00
7	7	1	92.86
7	11	0	100.00
6	10	0	100.00
7	9	0	100.00
7	17	0	100.00
5	9	0	100.00
5	5	0	100.00
7	12	0	100.00
7	18	0	100.00
12	11	0	100.00
7	14	0	100.00
7	10	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
4	14	1	94.44
3	13	0	100.00
13	11	0	100.00
7	8	1	93.33
5	14	0	100.00
6	14	0	100.00
8	11	0	100.00
6	9	0	100.00
7	11	1	94.44
10	11	0	100.00
3	14	0	100.00
5	17	0	100.00
6	15	0	100.00
9	18	0	100.00
7	6	0	100.00
9	12	0	100.00
6	18	0	100.00
7	18	0	100.00
4	12	0	100.00
7	11	0	100.00
8	11	0	100.00
6	10	0	100.00
3	17	0	100.00
5	14	0	100.00
7	12	0	100.00
6	16	0	100.00
7	14	0	100.00
6	9	0	100.00
7	17	0	100.00
5	23	0	100.00
7	13	0	100.00

Number of Observations Excluded		Number Misclassified	% Correctly Classified
Europe	NA 69		
9	14	1	95.65
7	20	0	100.00
4	9	0	100.00
5	11	0	100.00
5	15	0	100.00
3	10	0	100.00
6	13	0	100.00
10	12	0	100.00
7	11	0	100.00
10	10	0	100.00
5	6	0	100.00
7	19	0	100.00
5	14	0	100.00
8	6	0	100.00
7	16	0	100.00
8	14	0	100.00
9	14	0	100.00
13	10	0	100.00
8	12	0	100.00
9	13	0	100.00
4	12	0	100.00
9	14	0	100.00
7	15	0	100.00
11	16	0	100.00
4	14	0	100.00
7	20	0	100.00
11	8	0	100.00
7	12	0	100.00
7	10	0	100.00
10	15	0	100.00
6	13	0	100.00

	Number of Observations Excluded		Number Misclassified	% Correctly Classified
	Europe	NA 69		
	5	12	0	100.00
	5	13	0	100.00
	6	12	0	100.00
	7	7	0	100.00
	7	14	0	100.00
	8	14	0	100.00
Total	3526	6465	32	99.68%

APPENDIX C

JACKKNIFE OF 5 EUROPEAN RIVERS

All variables entered with covariate STNDLEN

n = 495
400 samples
p = 0.9494

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
9	6	2	5	4	3	88.46
6	9	2	5	2	3	87.50
5	7	4	8	5	0	100.00
6	11	1	1	3	2	90.91
6	3	5	4	1	1	94.74
3	7	2	7	2	1	95.24
9	4	4	4	5	2	92.31
6	7	2	3	4	4	81.82
9	12	8	6	2	1	97.30
2	8	2	5	8	1	96.00
8	9	3	2	5	0	100.00
6	7	3	6	5	2	92.59
6	11	3	9	1	1	96.67
9	6	4	2	1	1	95.45
7	4	4	5	0	1	95.00
4	7	4	8	4	0	100.00
8	7	4	5	3	1	96.30
2	7	3	2	4	2	88.89
8	15	4	5	3	3	91.43
8	9	2	5	4	1	96.43
6	6	4	5	3	1	95.83
5	8	5	7	3	4	85.71
3	4	2	4	5	1	94.44
7	8	1	5	3	0	100.00
4	6	2	7	3	0	100.00
4	9	2	7	5	2	92.59
7	15	2	4	2	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
5	9	1	2	0	0	100.00
4	1	6	7	3	0	100.00
6	9	2	12	5	1	97.06
1	7	3	5	4	0	100.00
3	9	4	5	3	0	100.00
6	4	2	6	3	2	90.48
7	7	0	10	6	2	93.33
12	8	0	4	0	1	95.83
7	8	1	3	3	1	95.45
4	10	1	3	2	2	90.00
5	4	1	8	2	1	95.00
7	12	4	4	5	0	100.00
7	10	0	3	5	1	96.00
9	7	0	2	5	1	95.65
9	7	4	8	3	1	96.77
4	11	1	8	4	3	89.29
7	12	5	7	6	0	100.00
3	10	4	2	4	2	91.30
6	7	1	6	2	0	100.00
5	10	2	4	6	0	100.00
7	7	4	4	2	2	91.67
7	5	4	5	5	1	96.15
6	5	3	4	4	1	95.45
13	10	3	7	5	2	94.74
2	8	4	8	1	1	95.65
4	14	2	8	5	3	90.91
5	8	3	5	6	0	100.00
6	10	4	2	1	1	95.65
0	5	0	9	5	0	100.00
6	7	0	4	6	3	86.96
7	9	2	9	6	0	100.00
5	4	4	7	1	1	95.24

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
7	12	1	8	2	1	96.67
3	4	1	3	1	0	100.00
7	8	2	11	1	0	100.00
6	8	1	7	2	1	95.83
4	7	0	6	4	1	95.24
4	13	3	4	6	0	100.00
6	10	4	8	4	1	96.88
3	8	1	7	8	0	100.00
5	6	3	5	1	0	100.00
5	6	5	4	3	3	86.96
8	7	1	4	4	2	91.67
13	8	4	6	3	1	97.06
6	9	5	3	3	1	96.15
6	8	3	5	2	2	91.67
2	10	6	7	6	2	93.55
1	6	2	4	3	0	100.00
6	8	1	3	7	1	96.00
4	5	0	7	5	1	95.24
3	12	1	1	7	0	100.00
6	5	2	8	4	1	96.00
6	7	2	3	1	0	100.00
6	5	1	7	4	1	95.65
4	12	6	5	2	0	100.00
3	6	1	4	4	0	100.00
8	8	2	10	2	0	100.00
6	4	4	5	1	1	95.00
7	11	5	7	3	1	96.97
5	12	1	3	1	0	100.00
4	10	4	10	6	2	94.12
4	5	5	5	4	1	95.65
6	8	3	7	3	5	81.48
6	8	2	9	3	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
3	7	3	5	4	0	100.00
6	6	3	7	3	1	96.00
9	8	4	4	0	1	96.00
4	6	1	3	5	0	100.00
6	8	4	9	5	1	96.88
3	8	2	9	3	2	92.00
8	10	2	7	1	1	96.43
4	10	1	3	1	0	100.00
2	11	6	5	2	0	100.00
9	10	3	10	9	1	97.56
4	8	4	3	4	0	100.00
6	8	3	7	6	0	100.00
6	9	2	8	5	0	100.00
8	6	5	2	4	1	96.00
5	8	0	2	3	0	100.00
7	10	2	6	5	0	100.00
6	8	2	6	4	1	96.15
7	6	2	5	1	0	100.00
10	9	2	2	3	0	100.00
8	9	0	4	1	1	95.45
10	8	1	2	9	1	96.67
6	11	6	5	3	0	100.00
4	6	2	3	4	0	100.00
7	6	0	4	7	0	100.00
5	6	4	3	2	0	100.00
4	10	1	4	5	1	95.83
8	9	3	5	3	3	89.29
5	6	4	6	5	0	100.00
4	4	2	6	6	2	90.91
10	7	6	4	2	1	96.55
4	6	4	4	8	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
4	12	0	4	3	0	100.00
6	6	0	2	3	0	100.00
3	10	1	5	2	1	95.24
5	0	4	8	2	1	94.74
7	8	2	6	3	1	96.15
7	6	4	9	6	2	93.75
9	7	1	5	4	4	84.62
3	11	2	5	7	2	92.86
2	10	5	3	3	0	100.00
6	13	5	7	2	3	90.91
8	9	2	5	8	0	100.00
5	9	4	2	4	1	95.83
6	6	3	3	8	0	100.00
3	5	8	6	0	0	100.00
6	9	1	6	6	4	85.71
6	9	2	7	5	2	92.86
8	6	5	4	9	1	96.88
4	5	1	3	2	2	86.67
5	9	2	3	2	0	100.00
2	8	3	4	6	1	95.65
8	7	2	5	2	1	95.83
5	18	6	4	3	0	100.00
3	4	5	6	2	2	90.00
2	10	1	10	4	2	92.59
7	11	1	7	0	0	100.00
5	6	2	6	2	0	100.00
6	12	1	5	4	3	89.29
4	7	1	9	9	0	100.00
4	5	1	7	2	1	94.74
0	11	1	5	6	0	100.00
4	5	4	2	5	2	90.00
6	11	1	6	3	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
5	6	2	4	4	1	95.24
9	6	3	3	5	0	100.00
6	10	4	5	5	2	93.33
7	11	3	5	4	2	93.33
7	16	5	4	3	2	91.43
6	8	0	9	4	1	96.30
2	5	1	7	2	1	94.12
4	9	1	9	5	1	96.43
2	11	3	6	5	2	92.59
9	8	3	6	2	0	100.00
3	9	2	4	5	0	100.00
2	6	3	8	2	0	100.00
3	6	4	3	3	3	84.21
3	6	3	4	7	1	95.65
11	12	1	8	2	2	94.12
8	7	2	5	3	0	100.00
7	11	4	3	1	0	100.00
3	7	3	2	1	0	100.00
4	13	2	2	5	1	96.15
7	7	4	3	3	0	100.00
1	8	1	6	3	1	94.74
6	8	2	5	5	2	92.31
7	7	3	7	3	3	88.89
5	9	0	6	5	1	96.00
5	10	1	9	4	1	96.55
1	2	2	2	7	1	92.86
8	9	3	6	6	1	96.88
4	10	3	6	4	3	88.89
6	5	3	3	4	2	90.48
5	15	2	5	3	1	96.67
11	20	2	10	4	3	93.62
3	14	4	4	4	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
10	4	1	6	6	1	96.30
5	12	3	1	2	3	86.96
7	10	1	7	1	1	96.15
4	7	3	5	9	3	89.29
12	5	2	6	5	0	100.00
1	8	2	9	6	2	92.31
7	7	4	5	3	2	92.31
4	11	4	2	5	2	92.31
7	12	1	6	2	0	100.00
7	6	4	7	5	2	93.10
5	4	2	4	2	3	82.35
7	4	0	3	2	0	100.00
3	6	6	2	6	1	95.65
4	7	3	5	5	0	100.00
4	4	3	2	4	0	100.00
12	4	8	8	3	2	94.29
5	9	2	5	6	2	92.59
6	5	4	2	4	1	95.24
6	8	1	3	3	1	95.24
10	14	2	4	1	2	93.55
8	7	3	7	6	1	96.77
7	9	2	6	1	2	92.00
6	11	2	8	4	0	100.00
4	9	5	6	4	0	100.00
6	7	3	1	3	3	85.00
7	11	3	7	2	1	96.67
3	4	4	2	4	1	94.12
7	5	1	2	5	4	80.00
8	6	2	4	4	0	100.00
5	2	0	2	5	1	92.86
8	5	3	6	2	0	100.00
8	6	3	8	8	4	87.88

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
6	7	4	6	3	1	96.15
10	9	5	5	3	0	100.00
6	4	1	4	6	3	85.71
4	9	4	4	5	4	84.62
4	5	3	3	5	1	95.00
11	6	1	4	3	2	92.00
2	5	6	3	4	0	100.00
5	9	2	7	4	2	92.59
5	7	4	2	2	0	100.00
5	6	2	5	4	0	100.00
5	11	5	2	6	1	96.55
6	11	4	3	3	0	100.00
8	8	0	5	4	1	96.00
6	6	5	6	4	2	92.59
4	3	1	4	3	1	93.33
3	9	4	4	5	1	96.00
2	9	2	3	3	2	89.47
8	3	2	7	3	0	100.00
7	10	3	2	3	1	96.00
11	5	2	6	5	5	82.76
6	12	4	0	2	3	87.50
8	4	6	5	1	0	100.00
6	6	3	4	7	1	96.15
5	10	2	4	4	1	96.00
13	6	5	5	3	0	100.00
3	5	3	4	4	1	94.74
5	11	2	3	4	4	84.00
5	9	6	1	4	2	92.00
8	10	3	3	7	1	96.77
9	10	0	4	3	1	96.15
4	8	1	5	3	1	95.24
3	8	3	6	3	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
4	5	2	6	2	0	100.00
2	10	3	4	6	0	100.00
4	11	5	2	4	1	96.15
5	7	2	8	4	1	96.15
6	7	0	6	5	2	91.67
4	6	3	7	5	2	92.00
4	6	3	8	3	1	95.83
5	10	0	7	2	1	95.83
6	8	3	4	2	1	95.65
4	10	2	4	2	2	90.91
5	5	2	4	5	0	100.00
5	8	2	6	5	1	96.15
3	3	0	4	6	0	100.00
1	7	4	4	1	2	88.24
4	7	2	4	5	2	90.91
4	6	4	7	6	0	100.00
9	4	0	4	3	2	90.00
5	4	4	7	1	2	90.48
8	12	5	1	2	2	92.86
7	9	1	3	6	0	100.00
6	13	3	4	5	2	93.55
5	3	5	8	1	0	100.00
11	8	1	1	4	2	92.00
9	13	0	5	7	3	91.18
6	11	2	7	1	0	100.00
5	9	2	4	3	1	95.65
7	13	2	4	5	1	96.77
6	9	5	4	3	1	96.30
7	13	2	5	5	1	96.88
6	6	4	6	1	2	91.30
6	5	1	7	7	1	96.15
8	7	3	7	3	2	92.86

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
5	8	4	5	5	1	96.30
5	9	6	4	2	0	100.00
5	5	4	4	3	0	100.00
9	3	4	4	4	2	91.67
0	4	1	10	4	2	89.47
3	4	4	6	3	2	90.00
8	12	4	6	4	1	97.06
4	4	3	5	5	2	90.48
5	11	3	5	4	1	96.43
6	11	3	4	5	2	93.10
8	7	2	7	2	0	100.00
7	9	5	2	3	0	100.00
6	4	2	8	3	1	95.65
5	4	1	3	4	1	94.12
5	13	2	8	5	1	96.97
7	10	3	6	5	3	90.32
6	9	0	2	5	1	95.45
7	4	1	3	2	1	94.12
6	8	5	6	5	1	96.67
4	9	2	2	2	1	94.74
4	3	3	7	2	1	94.74
7	7	2	2	4	2	90.91
1	6	3	8	6	2	91.67
6	3	5	2	5	1	95.24
5	3	1	8	9	4	84.62
5	10	1	7	5	0	100.00
1	9	0	6	1	0	100.00
8	9	4	6	8	3	91.43
4	11	2	5	4	1	96.15
7	6	3	6	4	1	96.15
8	9	5	6	1	1	96.55
7	7	3	4	2	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
5	10	2	3	3	3	86.96
6	5	3	7	3	1	95.83
10	7	3	6	4	1	96.67
3	7	3	4	3	0	100.00
7	7	2	5	5	3	88.46
7	7	4	5	4	2	92.59
5	8	5	7	3	2	92.86
4	9	1	7	2	3	86.96
5	9	3	5	0	1	95.45
9	6	4	2	5	2	92.31
5	5	4	2	1	0	100.00
8	13	1	8	2	2	93.75
9	7	1	9	5	3	90.32
9	8	0	7	7	3	90.32
6	8	4	5	3	1	96.15
7	5	3	5	6	1	96.15
6	8	3	3	3	2	91.30
10	6	2	6	3	5	81.48
10	5	2	1	6	1	95.83
2	9	4	6	2	1	95.65
2	8	4	5	1	3	85.00
2	7	3	4	1	0	100.00
8	11	1	3	1	2	91.67
5	9	3	5	0	1	95.45
10	13	1	7	3	0	100.00
6	10	2	7	5	3	90.00
9	7	0	1	4	2	90.48
13	5	2	6	3	2	93.10
5	2	4	5	7	1	95.65
6	10	4	5	2	1	96.30
7	4	6	4	6	2	92.59
2	6	1	2	2	0	100.00

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
1	7	1	2	4	0	100.00
7	7	2	4	4	1	95.83
8	4	3	9	3	0	100.00
7	12	2	9	4	1	97.06
6	13	3	5	2	0	100.00
9	8	0	4	3	2	91.67
9	8	2	4	6	2	93.10
10	8	2	6	5	0	100.00
10	3	6	3	5	1	96.30
4	5	1	5	4	1	94.74
6	7	2	2	3	0	100.00
4	6	4	6	3	2	91.30
10	12	1	4	1	1	96.43
7	7	3	11	1	0	100.00
6	9	2	6	3	0	100.00
6	9	6	9	4	1	97.06
9	9	3	4	4	2	93.10
9	14	4	5	6	2	94.74
3	7	7	2	3	0	100.00
4	10	5	8	3	0	100.00
3	7	1	1	6	1	94.44
3	6	4	6	2	1	95.24
6	12	3	9	7	0	100.00
4	5	0	6	3	2	88.89
6	6	2	5	1	0	100.00
8	11	4	3	8	2	94.12
8	13	3	4	3	2	93.55
2	7	3	4	3	0	100.00
4	5	2	8	3	3	86.36
8	11	1	5	3	2	92.86
9	11	2	3	2	1	96.30
4	11	5	9	2	1	96.77

Number of Observations Excluded					Number	% Correctly
Logan	Almond	Boyne	Lee	Usk	Misclassified	Classified
5	5	1	3	1	1	93.33
5	6	3	3	6	0	100.00
4	3	0	5	4	9	100.00
4	12	1	3	5	2	92.00
6	6	4	5	4	2	92.00
5	13	1	4	4	2	92.59
7	10	6	5	3	2	93.55
5	8	3	5	6	0	100.00
3	8	1	4	1	1	94.12
7	10	2	8	0	0	100.00
5	8	2	6	3	1	95.83
5	5	8	3	4	1	96.00
1	9	6	4	3	1	95.65
8	6	5	5	3	2	92.59
11	6	5	5	5	5	84.38
5	5	1	5	2	0	100.00
5	11	2	4	5	2	92.59
7	11	7	1	6	1	96.88
6	7	2	1	3	0	100.00
8	10	2	4	3	2	92.59
5	7	2	4	1	1	94.74
9	7	5	2	5	2	92.86
2316	3161	1081	2020	1484	470	95.34% Total

REFERENCES

1. Albrecht, G. H. (1978). Some comments on the use of ratios, Syst. Zool., 27, 67-71.

2. Anderson, T. W. (1958). An introduction to multivariate statistical analysis, New York: John Wiley and Sons.

3. Andrews,, D. F., Gnanadesikan, R., and Warner, J. L. (1971). Transformations of multivariate data, Biometrika, 27, 825-40.

4. Atchley, W. R. (1978). Ratios, regression intercepts, and the scaling of data, Syst. Zool., 27, 78-83.

5. Atchley, W. R., Gaskings, C. T., and Anderson, D. (1976). Statistical properties of ratios. I. Emperical results, Syst. Zool., 25, 137-48.

6. Bilton, H. T. (1971). Identification of major British Columbia and Alaska runs of even-year and odd-year pink salmon from scale characters. J. Fish. Res. Bd. Canada, 29, 295-301.

7. Blackith, R. E. and Reyment, R. A. (1971). Multivariate morphometrics, London and New York: Academic Press.

8. Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, J. Roy. Stat. Soc., (B), 26, 211-52.

9. Casselman, J. M., Collins, J. J., Crossman, E. J., Ihssen, P. E., Spangler, G. R. (1981). Lake whitefish (coregonus clupeaformis) stocks of the Ontario waters of Lake Huron. Can. J. Fish. Aquat. Sci., 38, 1772-89.

10. Chatfield, C. and Collins, A. J. (1980). Introduction to multivariate analysis, London and New York: Chapman and Hall.

11. Dillon, W. R. and Goldstein, M. (1984). Multivariate analysis, New York: John Wiley.

12. Dodson, P. (1978). On the use of ratios in growth studies, Syst. Zool., 27, 62-67.

13. Everitt, B. (1980). Cluster analysis, New York: Halsted Heinemann.

14. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, Ann. Eugen., 7, 179-88.

15. Geisser, S. (1964). Posterior odds for multivariate normal classifications, J. Roy. Stat. Soc., (B), 26, 69-76.

16. Gould, S. J. (1966). Allometry and size in ontogeny and phylogeny, Biol. Rev., 41, 587-640.

17. Johnson, R. A. and Wichern, D. W. (1982). Applied multivariate statistical analysis, London: Prentice-Hall.

18. Klecka, W. R. (1980). Discriminant analysis. Sage University paper series. Quantitative applications in the social sciences, series no. 07-019. Beverly Hills and London: Sage Publications.

19. Kenchington, T. J. (1986). Morphological comparison of two northwest Atlantic redfishes, Sebastes fasciatus and S. mentella, and techniques for their identification, Can. J. Fish. Aquat. Sci., 43, 781-87.

20. Kullback, S. (1959). Information theory and statistics, New York: John Wiley.

21. Lachenbruch, P. A. (1975). Discriminant analysis, New York: Macmillan.

22. Lachenbruch, P. and Mickey, R. M. (1968). Estimation of Error Rates in Discriminant Analysis. Technometrics, 10, 1-11.

23. Lear, W. H., and Misra, R. K. (1978). Clinal variation in scale characters of Atlantic salmon (Salmo salar) based on discriminant function analysis, J. Fish. Res. Board Can., 35, 43-47.

24. MacCrimmon, H. R. and Claytor, R. R. (1984). Meristic and morphometric identity of Baltic stocks of Atlantic salmon (salmo salar), Can. J. Zool., 63, 2032-37.

25. MacCrimmon, H. R. and Claytor, R. R. (1986). Possible use of taxonomic characters to identify Newfoundland and Scottish stocks of Atlantic salmon, salmo salar L., Aqua. Fish. Manag., 17, 1-17.

26. Misra, R. K. and Ni, I. H. (1983). Distinguishing beaked redfishes (deepwater redfish, Sebastes mentella and Labrador redfish, S. fasciatus) by discriminant analysis (with covariance) and multivariate analysis of covariance, Can. J. Fish. Aquat. Sci., 40, 1507-11.

27. Moore, P. G. and Tukey, J. W. (1954). Answer to query 112, Biometrics, 10, 562-68.

28. Morrison, D. F. (1976). Multivariate statistical methods, New York: McGraw-Hill.

29. Mosimann, J. E. and James, F. C. (1979). New statistical methods. Allometry with applications to Florida red-winged blackbirds, Evolution (Lawrence, Kansas), 33, 444-59.

30. Pimentel, R. A. (1979). Morphometrics. Kendal Hunt Co., Dubuque, IA.

31. Reddin, D. G. (1982). Some general information on discriminant functions and accuracy for identifying North American and European Atlantic salmon caught at West Greenland, ICES CM 1982/M:15, 15 pp.

32. Reddin, D. G. (1986). Discrimination between Atlantic salmon (Salmo salar L.) of North American and European origin. J. Cons. int. Explor. Mer., 43, 50-58.

33. Reist, J. D. (1985). An empirical evaluation of several univariate methods that adjust for size variation inmorphometric data. Can. J. Zool., 63, 1429-39.

34. Ritter, J. A., Marshall, T. A., Reddin, D. G. and Doubleday, W. G. (1980). Assessment of the impact of the West Greenland Atlantic salmon (Salmo salar) fishery on stocks and catches in North America, ICES CM 1980/M:38, 10 pp.

35. Shaklee, J. B. and Tamaru, C. S. (1981). Biochemical and morphological evolution of Hawaiian bonefishes (albula), Syst. Zool., 30, 125-46.

36. Sharp, J. C., Able, K. W., Leggett, W. C. and Carscadden, J. E. (1978). Utility of meristic and morphometric characters for identification of capelin (Mallotus villosus) stocks in Canadian Atlantic waters, J. Fish. Res. Board Can., 35, 124-30.

37. Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). Modern multivariate statistical analysis: a graduate course and handbook, Ohio: American Sciences Press.

38. Snedecor, G. W. and Cochran, W. G. (1967). Statistical methods, Iowa: Iowa State.

39. SPSS Inc. (1983). SPSS-X user's guide, New York: McGraw-Hill.

40. SPSS Inc. (1985). SPSS-X advanced user's guide, New York: McGraw-Hill.

41. Srivastava, M. S. and Carter, E. M. (1983). An introduction to applied multivariate statistics, New York: North-Holland.

42. Tatsuoka, M. M. (1971). Multivariate analysis, New York: John Wiley.

43. Velleman, P. F. and Hoaglin, D. C. (1981). Applicaitons, basics and computing of exploratory data analysis, Boston: Duxbury Press.

44. Wilk, S. J., Smith, W. G., Ralph, D. E. and Sibunka, J. (1980). Population structure of summer flounder between New York and Florida based on linear discriminant analysis, Trans. Am. Fish. Soc., 109, 265-71.