UNITED STATES DEPARTMENT OF THE INTERIOR, Fred A. Seaton, *Secretary*

FISH AND WILDLIFE SERVICE, Arnie J. Suomela, *Commissioner*

SOME USES OF STATISTICAL ANALYSIS IN CLASSIFYING RACES OF AMERICAN SHAD (*Alosa sapidissima*)

BY DONALD R. HILL

FISHERY BULLETIN 147

From Fishery Bulletin of the Fish and Wildlife Service

VOLUME 59

UNITED STATES GOVERNMENT PRINTING OFFICE • WASHINGTON : 1959

For sale by the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C.
Price 20 cents

Library of Congress catalog card for this bulletin:

Hill, Donald R

Some uses of statistical analysis in classifying races of American shad (*Alosa sapidissima*) Washington, U. S. Govt. Print. Off., 1959.

iv, 269–286 p. tables. 26 cm. (U. S. Fish and Wildlife Service. Fishery bulletin 147)

"From Fishery bulletin of the Fish and Wildlife Service, volume 59."
Bibliography: p. 283.

1. Shad. I. Title: Statistical analysis in classifying races of American shad. (Series: U. S. Fish and Wildlife Service. Fishery bulletin, v. 59, no. 147)

U. S. Dept. of the Interior. Library
for Library of Congress

Library of Congress catalog card for the series, Fishery Bulletin of the Fish and Wildlife Service:

U. S. *Fish and Wildlife Service.*

Fishery bulletin. v. 1–
Washington, U. S. Govt. Print. Off., 1881–19

v. in illus., maps (part fold.) 23–28 cm.

Some vols. issued in the congressional series as Senate or House documents.
Bulletins composing v. 47– also numbered 1–
Title varies: v. 1–49, Bulletin.
Vols. 1–49 issued by Bureau of Fisheries (called Fish Commission, v. 1–23)

1. Fisheries—U. S. 2. Fish-culture—U. S. I. Title.

Library of Congress [r55e¾]

CONTENTS

ABSTRACT

Each year pound nets fished in the ocean off the coasts of New York and New Jersey catch large quantities of shad. The majority of these fish are believed to be native to the Hudson and Connecticut Rivers and therefore, these catches should be considered in any management plan for the two rivers. To establish a management plan which would include the ocean fisheries, estimates of the racial composition of this catch must be made.

In this paper an analysis of some meristic counts for shad is presented to support the racial theory. Samples were examined and it was found that the meristic counts used could be considered representative of the populations. An analysis of variance of the characters gave evidence for the existence of races.

A discriminant function is presented whereby a mixed sample of Hudson and Connecticut River shad can be separated. Meristic data collected from Hudson River shad in 1939 and Connecticut River shad in 1945 are used to construct the discriminant function. The mean value of this function for the Hudson River, 1939, is 74.103 and for the Connecticut River, 1945, is 70.940.

The discriminant function obtained will correctly classify approximately 81 percent of a mixed sample of Hudson and Connecticut River shad. Meristic data collected from the Hudson River in 1940 were substituted into this discriminant function and out of 105 fish, 16 were incorrectly classified; this is in good agreement with the theoretical 19 percent misclassification. The number of misclassifications can be considerably reduced if the individuals falling close to the mid-point between the two populations are not classified. By refusing to classify about one-half of the sample, the number of wrong classifications is reduced to 3.7 percent. Several methods of estimating the population composition of a mixed sample of shad are presented.

SOME USES OF STATISTICAL ANALYSIS IN CLASSIFYING RACES OF AMERICAN SHAD (*Alosa sapidissima*)

By Donald R. Hill, Fishery Research Biologist, Bureau of Commercial Fisheries

The commercial catch of American shad (*Alosa sapidissima*) has declined since the beginning of the twentieth century. In 1950 a study of this species was undertaken by the Fish and Wildlife Service acting as the primary research agency of the Atlantic States Marine Fisheries Commission. The objectives of the investigation were to determine the causes for the decline in shad abundance, determine conditions favoring recovery, and provide basic information so that the fishery can be managed to obtain optimum yields.

Most American shad landed on the Atlantic coast are captured in rivers; however, pound nets fished off the coasts of New York and New Jersey take large numbers of them each spring. The racial origin of these fish must be known for the intelligent management of the species. In this study, use is made of meristic counts for shad to separate two races or populations.

Meristic data were collected under the supervision of Louella E. Cable of the U. S. Fish and Wildlife Service. Grateful acknowledgement is made to G. B. Talbot, Chief, Middle Atlantic Fishery Investigations, for supplying these data and reviewing the manuscript, to C. H. Walburg in preparing the manuscript for publication, to T. M. Widrig and D. D. Worlund for suggestions concerning the estimation of relative abundance; and to T. A. Bancroft of the Statistical Laboratory, and K. D. Carlander of the Zoology Department, Iowa State College, for numerous suggestions.

STATEMENT OF THE PROBLEM

Studies of the shad populations of the Hudson (Talbot 1954) and Connecticut (Fredin 1954) Rivers have shown the effect of fishing effort on catches made in subsequent years. As a result of analysis of catch-and-effort statistics and tagging experiments, the size of the runs in previous years was determined for each river. Table 1 shows the size of the catches for the two rivers for 1938–51 and the calculated fishing rates for each of these years. These catches include only the fish taken in the rivers and not fish caught in the ocean.

TABLE 1.—*Total shad catches and estimated fishing rates for the Hudson and Connecticut Rivers from 1938 to 1951*

Year	Hudson River		Connecticut River	
	Catch (1,000 pounds)	Fishing rate (percent)	Catch [1] (1,000 pounds)	Fishing rate (percent)
1938	2, 417	74. 3	376	46. 1
1939	3, 103	69. 9	332	43. 4
1940	3, 036	67. 2	278	33. 5
1941	3, 112	68. 4	364	32. 8
1942	3, 164	68. 3	344	34. 2
1943	3, 185	71. 0	478	44. 8
1944	4, 175	76. 3	636	56. 1
1945	3, 545	64. 7	651	70. 3
1946	3, 274	78. 6	899	81. 9
1947	2, 046	79. 0	657	81. 0
1948	2, 461	76. 3	532	73. 5
1949	2, 038	74. 3	392	69. 6
1950	992	70. 9	231	58. 8
1951	755	46. 0	303	56. 7

[1] Catch for Connecticut River estimated by a factor of 3 pounds per fish.

Through an analysis of data on scales of 6-year-old shad from the Connecticut River, Fredin (1954) found for shad an extraneous mortality of about 40 percent occurring outside the river fishery. This was nearly as great as the fishing mortality in the river. He suggested that the pound nets in the New York Bay area and along the New Jersey coast could be the cause of some of this mortality. The number of pound nets in operation increased from 144 in 1946 to 180 in 1950. For these same years, the estimated populations (estimated by a regression analysis of escapements in previous years) were higher than the populations calculated from the catch-and-effort data. Fredin stated that the increase in pound-net effort may account for these deviations.

NOTE.—Approved for publication December 4, 1957. Fishery Bulletin 147.

He made the following statement about these pound-net catches: (p. 258)

> The relation between pound-net catches and deviations from the expected populations in the Connecticut River cannot be fully evaluated at this time because the extent to which the Connecticut River shad contribute to these pound-net catches is not known. Additional tagging studies conducted in the areas where pound nets are fished would enable us to determine the effect of this fishing on the Connecticut River shad runs. The causes of the extraneous-mortality rate must be taken into consideration in a management program to restore the Connecticut River shad population to the level of abundance which it held in the early 1940's.

The extent of the shad fisheries in New York Bay and along the Long Island and New Jersey coasts can be seen from the 1945 catches of shad reported by the U. S. Fish and Wildlife Service (1949). Total catches in pounds, by county, are given south to north from southern New Jersey to Long Island.

Atlantic Co., N. J.	60, 700
Ocean Co., N. J.	690, 900
Monmouth Co., N. J.	1, 173, 600
Suffolk Co., L. I., N. Y.	217, 000

This total New Jersey coast, Long Island, and New York Bay catch is about two million pounds while the combined total catch for the Hudson and Connecticut Rivers is about 4.2 million pounds for the same year. Of course the composition of this New Jersey catch is the basic problem. If it is primarily shad from southern rivers, this catch can be disregarded in the management of the Hudson and Connecticut Rivers. Conversely, if this catch is predominantly Hudson and Connecticut River shad, it must be considered in any management program because it represents one-third of the total fishery.

The Fish and Wildlife Service has carried out some tagging experiments[1] in the areas under consideration. An examination of these tag returns can supply a partial solution to the composition of these three catches. In 1945, 125 shad were tagged at Seaside Park, N. J. The following areas and numbers of recaptures were reported:

Hudson River	20
Connecticut River	3
Delaware River	2
Chesapeake Bay	4
New York Ocean	2
New Jersey Coast	5
Maine Coast	1

In the same year 97 shad were tagged off Fire Island Inlet, Long Island, N. Y., and were recovered in the following areas:

Hudson River	9
Connecticut River	24
Bay of Fundy	1

Shad have been tagged off Staten Island, New York Bay, in several different years, and the following table gives the recoveries from these experiments in which a total of 1,380 shad were tagged.

Hudson River	448
Connecticut River	24
Delaware River	5
Chesapeake Bay	1
Long Island	5
New Jersey Coast	13
Bay of Fundy	3

These tagging experiments furnish us with some information about the composition of the populations in these three areas. Most of the shad tagged on the New Jersey coast migrated into the Hudson River. Similarly, a major part of those tagged off Staten Island were recaptured in the Hudson River. On the coast of Long Island, most of the fish tagged were recaptured in the Connecticut River. Very few of the fish tagged in these areas were recaptured in other major shad rivers.

The tag returns could be used in conjunction with the catch statistics of the various areas to estimate the composition of the catches. They have been of considerable value in showing the general composition of the catches in the areas under question, and it is apparent that these catches should be included in any analysis of the catch and effort statistics for either river. Any increase or decrease in effort in these areas will be reflected in the number of shad entering the rivers.

Unfortunately, complete catch and effort statistics are not available for the New Jersey pound-net fishery, so it is impossible to compare directly past pound-net catches and mortality rates of Hudson or Connecticut River shad. However, a research project could be designed to show what effect pound-net catches have on these mortality rates. This could be done by dividing the pound-net fishing areas into a number of geographical strata. For each stratum the total catch and effort would be needed. It would also be necessary to estimate the composition of this catch (for

[1] Unpublished data, U. S. Fishery Laboratory, Beaufort, N. C.

example, 70 percent Hudson shad, 30 percent Connecticut shad) for each stratum. From these quantities, the entire pound-net catch could be divided into two parts, Hudson River shad and Connecticut River shad.

Obtaining the total catch and effort statistics is rather straightforward and not expensive, but estimating the compositions of the catches is much more complex. At the present time, this would have to be done by tagging experiments. In each stratum, a number of shad would be tagged, preferably at various times throughout the season. If the fishermen are personally interviewed to obtain tag returns, the coverage on the two rivers should be equal. Even the fishing rates for the two rivers should be the same or some adjustment would be needed to place them on the same level. As a result, the problem of estimating the composition of a pound-net catch becomes complex and expensive when two river systems are canvassed for tag returns, and the tagging is done several hundred miles from the rivers.

This is one of many fishery problems where it would be advantageous to obtain a sample of fish and classify them according to the river system to which they belong. If this were done accurately, samples of fish from various strata could be obtained, and each fish could be assigned to the proper river. The composition of this sample would be used to estimate the composition of the stratum. The remainder of this paper will investigate the statistical techniques applicable to this problem.

SOURCE OF MERISTIC DATA

Races (populations) of fish can often be separated through the use of body measurements or meristic counts. If some of these counts or measurements are sufficiently different for two populations, it is possible to classify the individuals in a sample and estimate the relative abundance of each population in an area by the composition of the sample.

A large number of morphological data were collected on the American shad, both juvenile and adults, by the Fish and Wildlife Service between 1939 and 1945. Data on twenty-five different characters were collected from each fish. They were defined as follows:

MID-CAUDAL LENGTH.—Tip of snout to end of shortest rays between lobes of caudal fin.

TOTAL LENGTH.—Tip of snout to end of longest ray of caudal fin.

STANDARD LENGTH.—Tip of snout to branching of urostyle (modified vertebra).

DEPTH.—Longest measurement from dorsal to ventral profiles (in front of dorsal fin).

THICKNESS OF FISH.—Measurement from left to right through thickest part of fish.

CAUDAL PEDUNCLE.—Shortest dorsoventral measurement of tail anterior to caudal fin.

HEAD LENGTH.—Tip of snout to posterior margin of opercular bone.

SNOUT.—Tip of snout to anterior margin of eye socket.

EYE.—From anterior to posterior margin of eye socket.

INTERORBITAL.—Across top of head from dorsal margin of one eye socket to dorsal margin of the other eye socket above pupil of the eye.

MAXILLARY.—From posterior margin of maxillary to a vertical from tip of snout.

LENGTH OF DORSAL AND ANAL BASES.—From anterior margin of base of first ray to posterior margin of last ray.

LENGTH OF PECTORAL.—From articulation of first ray to tip of longest ray.

SNOUT TO DORSAL.—Tip of snout to articulation of first ray of dorsal fin.

SNOUT TO ANAL.—Tip of snout to articulation of first ray of anal fin.

PECTORAL TO VENTRAL.—From articulation of first ray of pectoral fin to articulation of first ray of ventral fin.

VENTRAL TO ANAL.—From articulation of first ray of ventral fin to articulation of first ray of anal fin.

ANTERIOR SCUTES.—All scutes having processes in front of ventral fins, including the scute between the fins which does not appear to have a process. It is beneath the process of the preceding scute.

POSTERIOR SCUTES.—All scutes posterior to ventral fins.

VERTEBRAE.—Urostyle included in count.

DORSAL RAYS.—Last undivided ray counted with divided rays, other undivided rays separate.

ANAL RAYS.—As for dorsal rays.

PECTORAL RAYS.—All rays on left and right sides of fish.

GILL RAKERS.—Only those on the lower limb of the first gill arch counted (at the bend of the arch, the bases of the rakers of the upper arm point in the opposite direction from those of the lower arm).

SCALES.—Oblique rows from the upper end of opercular slit to base of caudal fin. Horizontal rows from the median dorsal line to ventral scutes but not including either.

Since there were numerous rivers involved and samples were taken for several years from some rivers, analysis using all the data would become exceedingly complex. Table 2 gives the location, year, and number of adult specimens examined.

TABLE 2.—*Summary of areas from which meristic data were collected on adult American shad*

[The figures represent the number of shad in each sample. Data were not complete for all meristic characters in all samples]

Area	1938	1939	1940	1941	1942	1943	1944	1945
Connecticut River	------	------	------	------	------	------	------	101
Hudson River (N. Y.)	------	104	105	102	------	------	------	------
Maurice River (N. J.)	------	------	------	------	------	------	------	100
Delaware River	------	------	------	------	------	------	68	------
Chesapeake Bay	105	100	------	102	------	------	------	------
Albemarle Sound (N. C.)	127	124	85	------	------	------	------	------
Edisto River (S. C.)	50	99	------	99	96	------	------	------
Ogeechee River (Ga.)	------	50	------	------	------	------	------	------
St. Johns River (Fla.)	45	100	106	------	------	------	------	------

This paper will not include a complete analysis of the data available. It is hoped that this preliminary analysis will show that a complete analysis would be warranted, and that further research along these lines would be fruitful. No analysis of the information on the juveniles has been attempted, hence there is still much to be learned by combining this with the data on the adults.

REPRESENTATIVENESS OF THE SAMPLES

The need for separation of races of fish is apparent. However, before races can be separated, it should be established that they are present. Tagging experiments have been offered as the best evidence supporting the theory of a separate race of shad in each Atlantic coast river. In numerous tagging programs carried out by the Fish and Wildlife Service, few tagged shad have been recaptured in rivers other than the one in which they were tagged. No shad tagged on the spawning ground of one river system has ever been recaptured on the spawning ground of another river system. The operation of the homing instinct may not be 100 percent for shad, but examination of tag returns indicates that the percentage is very high.

If there is a race of shad in each river, that is, a group of fish and their offspring which return to the same spawning area year after year, the fish within a river should be more like one another than to the shad from other rivers. This could be expected because of environmental differences or genetic isolation. Conversely, if the spawning ground of each fish is determined by a completely random process, a single homogeneous shad population would be expected. Therefore, if consistent differences between shad in the several rivers can be found for some measurable characters, this can be used as further evidence to support the race theory.

Before proceeding with an attempt to verify this theory, one assumption should be investigated. If the available data are to be used to establish differences between rivers, it is essential that the samples be representative of the various populations. It is impossible to assume that these are random samples, because the shad fishermen know that they can control the size of fish in their catches by changing the mesh size of their nets. They know that if they fish a 5¾-inch stretched-mesh gill net, they will catch large-roe shad, and if they use a 4¾- to 5-inch stretched-mesh net, they will catch proportionately more small shad of both sexes. This selectivity occurs with drift, anchor, and stake gill nets. Pound nets and haul seines may be much less selective, but if the gill nets are catching large fish and permitting smaller ones to escape, the population being sampled by the haul seines is not the total population of the river, but the total population minus the fish removed by gill nets. The result, of course, would be an excess of smaller fish in the haul seine samples.

Since it is known that some fishing gear tends to select fish by length (and all correlated measurements such as depth and thickness), it cannot be assumed that the samples are random. However, this selectivity may take place only in the size of the fish in the samples and not in some of the other characteristics. If the number of rays in the pectoral fin is being investigated, the samples may be representative of this character even if there is selectivity of size. This would be true of any character which is not correlated with length. Therefore, the characters were tested for a correlation with length and if none was found, the samples were considered representative.

In examining the catches of shad in the St. Johns River of Florida and the Connecticut River, it is apparent that the shad in the Connecticut are larger. Some of this can probably be explained by the difference in the age distributions of the two populations. Can these age distributions be used in separating races of shad? The author feels that they cannot be used, since they will fluctuate from year to year with changes in fishing effort and catches. Age and length are correlated, and this is another reason for excluding length and all correlated characters from any racial investigation.

The choice of characters to be used in investigating races of shad was evident. To avoid the difficulties presented by selectivity of the fishing gear, those characters which are correlated with length were not considered. Thus, depth, thickness, weight and all of the other body measurements were eliminated, because they increase as the fish grows. Ratios of two such measurements will also be related to length, unless these two measurements increase at the same rate throughout the growth of the fish. These ratios have not been investigated because it is doubtful that this condition exists, particularly when both juveniles and adults are considered. Scattergrams of gill rakers and length exhibited a parabolic relation and were therefore eliminated. Scale counts have not been included because notations on the data sheets indicated that some of the scale counts were questionable. Of the 25 counts and measurements, all were eliminated for the above reasons except 6 meristic counts (anterior scutes, posterior scutes, dorsal rays, anal rays, pectoral rays, and vertebrae) and these were tested for correlations with length before they were used in any analysis.

Analysis of variance tables for the regression of these characters on length have been calculated for some of the samples. The assumptions necessary for this analysis are: (1) For each length, the character is normally distributed, (2) the variance of the character is homogeneous for each length, and (3) for each length interval, the samples are random. The values of F needed to test for a regression of the characters on length are given in the last column of table 3. F with 1 and 100 degrees of freedom is equal to 3.94 at the 5 percent level and 6.90 at the 1 percent level. There are three significant regressions in this table. Two of these regressions are for vertebrae in the Hudson River samples of 1939 and 1940. It is interesting to note that the F-value for vertebrae in the 1941 Hudson River sample is also high (3.11) but not significant. This significance does not occur in any of the other three samples which were tested for a regression of vertebrae on length. Of course, this is not enough evidence in itself to say that it is a racial difference between the populations, but it raises the question as to why this difference occurs for the Hudson River samples

TABLE 3.—*Regression analyses to test meristic characters for correlation with length*

Source of variation	df	Sum of squares	Mean square	F
ANTERIOR SCUTES				
HUDSON RIVER, 1940				
Regression	1	1.270	1.270	2.91
Deviation from regression	103	44.958	0.436	
Total	104	46.228		
ST. JOHNS RIVER, 1940				
Regression	1	1.241	1.241	2.33
Deviation from regression	104	55.297	0.532	
Total	105	56.538		
CONNECTICUT RIVER, 1945				
Regression	1	0.180	0.180	0.45
Deviation from regression	99	39.820	0.402	
Total	100	40.000		
POSTERIOR SCUTES				
HUDSON RIVER, 1940				
Regression	1	0.236	0.236	0.46
Deviation from regression	103	52.310	0.508	
Total	104	52.546		
ST. JOHNS RIVER, 1940				
Regression	1	0.024	0.024	0.44
Deviation from regression	104	57.176	0.550	
Total	105	57.200		
CONNECTICUT RIVER, 1945				
Regression	1	2.390	2.390	3.71
Deviation from regression	99	63.800	0.644	
Total	100	66.190		
ANAL RAYS				
HUDSON RIVER, 1940				
Regression	1	0.151	0.151	0.15
Deviation from regression	103	102.382	0.994	
Total	104	102.533		
ST. JOHNS RIVER, 1940				
Regression	1	0.002	0.002	0.002
Deviation from regression	104	106.762	1.027	
Total	105	106.764		
PECTORAL RAYS				
HUDSON RIVER, 1940				
Regression	1	0.644	0.644	1.64
Deviation from regression	103	40.346	0.392	
Total	104	40.990		
ST. JOHNS RIVER, 1940				
Regression	1	0.002	0.002	0.005
Deviation from regression	104	43.847	0.422	
Total	105	43.849		
CONNECTICUT RIVER, 1945				
Regression	1	1.425	1.425	2.15
Deviation from regression	98	65.075	0.664	
Total	99	66.500		
VERTEBRAE				
HUDSON RIVER, 1939				
Regression	1	2.606	2.606	*3.97
Deviation from regression	102	66.945	0.656	
Total	103	69.551		
HUDSON RIVER, 1940				
Regression	1	3.431	3.431	*4.41
Deviation from regression	103	80.089	0.778	
Total	104	83.520		
HUDSON RIVER, 1941				
Regression	1	2.386	2.386	3.11
Deviation from regression	96	73.746	0.768	
Total	97	76.132		

Footnote at end of table.

TABLE 3.—*Regression analyses to test meristic characters for correlation with length*—Continued

Source of variation	df	Sum of squares	Mean square	F
VERTEBRAE—Continued				
ST. JOHNS RIVER, 1938				
Regression	1	0.002	0.002	0.002
Deviation from regression	41	39.440	0.962	
Total	42	39.442		
ST. JOHNS RIVER, 1940				
Regression	1	0.115	0.115	0.17
Deviation from regression	104	69.942	0.672	
Total	105	70.057		
CONNECTICUT RIVER, 1945				
Regression	1	0.684	0.684	0.50
Deviation from regression	99	134.316	1.357	
Total	100	135.000		
DORSAL RAYS				
HUDSON RIVER, 1940				
Regression	1	0.001	0.001	0.002
Deviation from regression	103	58.532	0.568	
Total	104	58.533		
ST. JOHNS RIVER, 1940				
Regression	1	3.650	3.650	**6.71
Deviation from regression	104	56.580	0.544	
Total	105	60.230		

NOTE.—Asterisks denote significant.

and not for the St. Johns and Connecticut River samples. The other significant *F* is for the regression of dorsal rays in the St. Johns River sample of 1940.

This regression analysis of samples from the Connecticut River, Hudson River, and St. Johns River shows that none of the six characters has a consistent correlation with length. It is difficult to explain the regression of vertebrae on length for the Hudson River sample. This is also true for dorsal rays in the St. Johns River sample; however, with 19 regressions tested, three significant values is a small proportion. Since none of the characters is consistently correlated with length, the available samples were considered representative even though they are not random.

STATISTICAL EVIDENCE FOR THE EXISTENCE OF RACES

It has been shown that none of the six characters: anterior scutes, posterior scutes, dorsal rays, anal rays, pectoral rays and vertebrae, exhibits a consistent correlation with length; therefore, we can place a certain degree of confidence in treating the samples as representative. In the previous section it was pointed out that consistent differences between rivers for some measurable characters would support the racial theory. This can be studied by setting up the following mathematical model:

$$Y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk}$$

where Y_{ijk} is the character under study, μ is the general mean, α_i represents the contribution of the ith river, β_j is the contribution of the jth year (year caught), and $(\alpha\beta)_{ij}$ is an interaction of years and rivers. The ϵ_{ijk} is an error term. The α_i, β_j, $(\alpha\beta)_{ij}$, and ϵ_{ijk} are all assumed to be normally and independently distributed with means zero and variance σ^2_R, σ^2_Y, σ^2_{RY} and σ^2, respectively.

This model could be changed so that the jth classification stood for year class, but this necessitates knowing the ages of all the fish in the samples. A model of the latter type may have an advantage, since differences between year classes would not be averaged as they are with the above model. Unfortunately, the present data do not include ages, hence the model indicated will be used. This will in no way invalidate the conclusions, but grouping by year class might be a refinement that would prove valuable.

The racial theory can now be investigated more fully. Using the above model and suitable data, several hypotheses can be tested which may give added support to this theory. First of all, an interaction of years with rivers ($H_0:\sigma^2_{RY}=0$) can be tested, next a test of differences between years ($H_1:\sigma^2_Y=0$), and third, a test for differences between rivers ($H_2:\sigma^2_R=0$). If H_0 and H_1 can be accepted while H_2 is rejected, the conditions necessary to support the race theory are present.

These hypotheses and their relations to the present problem will be explained in some detail. The first one, ($H_0:\sigma^2_{RY}=0$), is a test for an interaction between rivers and years. This interaction could best be described by assuming that temperature is a factor in determining the number of vertebrae of young shad. If there were a warm spring on a northern river and a cold spring on a southern river in 1953, and just the opposite in 1954, there might be produced the following average number of vertebrae for shad from the two rivers:

Year	Northern river	Southern river
1953	55.7	56.3
1954	56.2	55.8

In this situation, one would conclude that vertebrae offer no evidence for the presence of races, and most of the variation is in the form of an interaction between rivers and years.

The second hypothesis to be tested, $(H_1:\sigma^2_Y=0)$, concerns a difference between years. If the average number of vertebrae for two rivers and two different years were of the following magnitude, they would offer no proof for the presence of races.

Year	River A	River B
1953	56. 5	57. 0
1954	57. 0	57. 5

In this case, the difference from year to year is as large as the difference between rivers.

The third hypothesis, $(H_2:\sigma^2_R=0)$, to be tested is the one for a difference between rivers. If the difference between rivers is not significant, there would be no evidence for the presence of races. Thus if H_0 and H_1 can be accepted and H_2 rejected, the conditions necessary for the presence of races would be satisfied.

Analysis of variance tables with years and rivers as the two classifications were computed for five of the six characters mentioned above. Dorsal rays had to be omitted because the data were incomplete. Table 2 shows that there are data for 1938 and 1939 in four locations: the St. Johns River in Florida, the Edisto River in South Carolina, Albemarle Sound, N. C., and Chesapeake Bay. Since the samples from Chesapeake Bay came from several different rivers, those data were not included in the analysis. The remaining three areas were used as one classification, and the years 1938 and 1939 were used as the other. A table of the same size could have been constructed using Hudson River, Albemarle Sound, and St. Johns data for the years 1939 and 1940.

The various samples are of unequal size ranging from 45 to 127 for the 2×3 table. Exact methods are available for the analysis of a 2×3 table with unequal subclass numbers but they require considerable computing time. Several approximations are available (Anderson and Bancroft, 1952) utilizing the complete data. In this study random samples of 40 fish each were drawn from the various samples, avoiding the difficulties of the unequal subclass numbers (table 10, appendix). With samples of size 40, this method should be a good approximation to the more exact methods.

TABLE 4.—*Analyses of variance for the meristic characters to test for differences between years, differences between rivers, and interaction between years and rivers*

Source of variation	df	Sum of squares	Mean square	F
ANTERIOR SCUTES				
Between years	1	1. 837	1. 837	3. 72
Between rivers	2	7. 599	3. 799	**7. 69
Y×*R*	2	0. 101	0. 051	0. 10
Error	234	115. 125	0. 494	
Total	239	124. 662		
POSTERIOR SCUTES				
Between years	1	2. 204	2. 204	3. 49
Between rivers	2	19. 733	9. 867	**15. 64
Y×*R*	2	2. 234	1. 117	1. 77
Error	234	147. 125	0. 629	
Total	239	171. 296		
ANAL RAYS				
Between years	1	0. 067	0. 067	0. 07
Between rivers	2	4. 059	2. 029	2. 17
Y×*R*	2	1. 408	0. 704	0. 76
Error	234	218. 200	0. 932	
Total	239	223. 734		
PECTORAL RAYS				
Between years	1	0. 416	0. 416	1. 20
Between rivers	2	11. 808	5. 904	**17. 01
Y×*R*	2	1. 859	0. 929	2. 68
Error	234	81. 100	0. 347	
Total	239	95. 183		
VERTEBRAE				
Between years	1	0. 066	0. 066	0. 09
Between rivers	2	19. 200	9. 600	**12. 87
Y×*R*	2	1. 634	0. 817	1. 10
Error	234	174. 500	0. 745	
Total	239	195. 400		

NOTE.—Asterisks denote significant.

Analysis of variance tables for these five characters are shown in table 4. The *F* values for testing the hypothesis of no interaction are the lower numbers in column five of this table. These range in value from 0.10 for anterior scutes to 2.68 for pectoral rays. None of these is significant ($F_{2,200}=3.04$ at the 5-percent level), so the hypothesis of no interaction of years and rivers is accepted.

The F values for testing differences between years range in value from 0.07 to 3.72. Again, these are not significant ($F_{1,200}=3.89$ at the 5-percent level), so the hypothesis of no differences between years can be accepted.

The F values for testing differences between rivers range in value from 2.17 to 17.01. The value for anal rays, 2.17, is not significant at the 5-percent level ($F_{2,200}=3.04$ at the 5-percent level, $F_{2,200}=4.71$ at the 1-percent level); however, the other four are all significant at the 1-percent level.

While the differences between rivers are not significant for anal rays, they are for the other four characters, so it can be safely concluded that there are differences from river to river for four characters.

The F values for differences between years are rather large for anterior scutes and posterior scutes, and both are significant at the 10-percent level ($F_{1,120}=2.75$). At the present time, it is impossible to say definitely whether there are differences from year to year for these two characters. It is apparent that this is one phase of the problem that should be studied in more detail.

The fact that there may be differences from year to year for some characters does not disprove the racial theory. The magnitude of the differences between years relative to the differences between rivers is the essential quantity to be considered in this problem. If the differences between years are small in comparison to the differences between rivers, races can still be distinguished. In terms of the model presented on page 274, the river-effects (α_i) should be considerably larger than the year-effects (β_j). An estimate of the relative magnitude of these two effects can be obtained from the analysis of variance tables. Since none of the interactions was significant, the interaction mean square has been pooled with the error mean square to obtain an estimate of the error (σ^2). It has been assumed that both the years and the rivers are a sample from a large number of years and rivers. Therefore, the α_i and β_j obtained from the data are samples from some larger population of α_i and β_j which have variance σ_R^2 and σ_Y^2. From the mean squares in the fourth column of table 5, estimates of σ_R^2 and σ_Y^2 can be obtained. For anterior scutes: $\hat{\sigma}^2=0.488$, $\hat{\sigma}_Y^2=0.0112$, and $\hat{\sigma}_R^2=0.0414$; for posterior scutes: $\hat{\sigma}^2=0.633$, $\hat{\sigma}_Y^2=0.0131$, and $\hat{\sigma}_R^2=0.1154$. Thus, the variation between years for anterior scutes is about one-fourth as large as the variation between rivers. Similarly, for posterior scutes, the variation between years is about one-ninth the variation between rivers.

These analyses present evidence that there are no differences between years for anal rays, pectoral rays, and vertebrae. There may be differences between years for anterior and posterior scutes, but if they do exist, they are small compared to the variation between rivers. These analyses of variance have approached the racial problem in a more direct manner than in previous studies and have given further support to the racial theory.

TABLE 5.—*Analysis of variance with interaction term pooled with error term*

Source of variation	df	Sum of squares	Mean square	Expected mean square
ANTERIOR SCUTES				
Between years	1	1.837	1.837	$\sigma^2+120\sigma_Y^2$
Between rivers	2	7.599	3.799	$\sigma^2+80\sigma_R^2$
Error	236	115.226	0.488	σ^2
Total	239	124.662		
POSTERIOR SCUTES				
Between years	1	2.204	2.204	$\sigma^2+120\sigma_Y^2$
Between rivers	2	19.733	9.867	$\sigma^2+80\sigma_R^2$
Error	236	149.359	0.633	σ^2
Total	239	171.296		

DISCRIMINANT FUNCTION ANALYSIS

There are numerous ways of using the data from meristic counts to construct discriminant functions. Raney and de Sylva (1953) constructed such a function by adding the number of dorsal, anal, and pectoral rays for each fish. They called this a "character index," but actually it is a simple form of a discriminant function. By plotting a frequency histogram of this character index for several areas and a series of years, they were able to differentiate to some extent between striped bass from the Hudson River and from Chesapeake Bay. There was considerable overlap in these distributions, and if one were presented with a fish of unknown origin, it would be difficult to assign it to a particular population with any certainty.

For illustration purposes, a discriminant function of this type has been constructed using data on shad from the Connecticut River and the Hudson River. It is unfortunate that the data for the Hudson River were collected in 1939 and 1940 (tables 11 and 12, appendix) while the Connecticut data were collected in 1945 (table 13, appendix), but since it has been shown that the characters are consistent from year to year, the data can be used for discrimination. The discriminant function,

$$Z=X_1+X_2+X_3+X_4+X_5+X_6$$

where X_1 is the number of anterior scutes, X_2 the number of posterior scutes, X_3 the number of dorsal rays, X_4 the number of anal rays, X_5 the number of pectoral rays, and X_6 is the number of

TABLE 6.—*Frequency distributions of the discriminant function*

$$Z=X_1+X_2+X_3+X_4+X_5+X_6$$

Z	Connecticut River 1945	Hudson River 1939	Hudson River 1940
139	1	----	----
140	----	----	----
141	1	1	----
142	3	----	----
143	4	1	----
144	14	----	1
145	10	2	2
146	18	3	6
147	10	6	3
148	10	11	8
149	9	18	18
150	6	18	15
151	4	18	23
152	1	12	9
153	----	8	12
154	----	5	3
155	----	----	2
156	----	1	----
157	----	----	2
158	----	----	1

vertebrae, has been tabulated for the Hudson River samples of 1939–40 and the Connecticut River sample of 1945 in table 6. It is interesting to note that the means of these distributions are: Hudson River (1939), 149.962 ($n=104$), Hudson River (1940), 150.362 ($n=105$), and the Connecticut River (1945), 146.363 ($n=91$); the variances are 5.816, 6.465, and 6.400, respectively. The pooled average for the Hudson River is 150.163. (A t-test shows that there is a highly significant difference between the two rivers.) If one were to use such a function for discrimination, he would classify everything above 148.2 as coming from the Hudson River and everything below as coming from the Connecticut River. However, since the counts are discrete, it would be necessary to use either 148 or 149 as the dividing line. Table 7 gives the percentage of wrong classifications for these two values. This simple function of the type $Z=\Sigma X_i$ provides a method of classifying about 78 percent of the individuals correctly. Without the use of this function, it would appear impossible to distinguish Connecticut River shad from Hudson River shad.

If such good results were obtained by totaling the number of scutes, vertebrae and rays for each specimen, perhaps some other combination might be more efficient. Considering only linear forms of the type $Y=\Sigma a_i X_i$, that function which is best for discriminating between the two populations can be determined. It can be shown (Rao 1952) that the best linear discriminant function for two multivariate normal populations is:

$$D=l_1X_1+l_2X_2+l_3X_3+l_4X_4+l_5X_5+l_6X_6$$

where the l_i's are obtained by solving the following set of equations:

$$\begin{aligned}
l_1w_{11}+l_2w_{12}+l_3w_{13}+l_4w_{14}+l_5w_{15}+l_6w_{16}&=d_1\\
l_1w_{21}+l_2w_{22}\quad\ldots\quad&=d_2\\
\cdot\qquad\qquad&\quad\cdot\\
\cdot\qquad\qquad&\quad\cdot\\
l_1w_{61}+l_2w_{62}\quad\ldots\quad l_6w_{66}&=d_6.
\end{aligned}$$

w_{ij} is an estimate of the covariance (assumed to be equal in the two populations) between the ith and jth characters and d_i is the estimated difference in mean values of the ith character in the two populations. The w_{ij} are estimated from the following equations:

$$(N_1+N_2-2)w_{ij}=\sum_{k=1}^{N_1}(X_{i1k}-\overline{X}_{i1})(X_{j1k}-\overline{X}_{j1})+\sum_{k=1}^{N_2}(X_{i2k}-\overline{X}_{i2})\ (X_{j2k}-\overline{X}_{j2}).$$

N_1 and N_2 are the number of specimens in the first and second sample, respectively, and X_{i1k} is the count on the ith character for the kth fish from population 1. $\overline{X}_{i1}$ is the mean value of the ith character for population 1.

Using data from the Hudson River sample of 1939 and the Connecticut River sample of 1945,[2] the following set of equations is obtained:

$$\begin{aligned}
&0.38197l_1+0.03742l_2+0.06242l_3-0.01515l_4+\\
&\qquad 0.02467l_5+0.15184l_6=0.41484\\
&0.03742l_1+0.71332l_2-0.02032l_3-0.01196l_4+\\
&\qquad 0.02301l_5+0.18071l_6=0.46016\\
&0.06242l_1-0.02032l_2+0.65354l_3+0.21084l_4+\\
&\qquad 0.03309l_5+0.09918l_6=0.71291\\
&-0.01515l_1-0.01196l_2+0.21084l_3+0.88481l_4+\\
&\qquad 0.00717l_5+0.13073l_6=0.38462\\
&0.02467l_1+0.02301l_2+0.03309l_3+0.00717l_4+\\
&\qquad 0.58499l_5-0.01020l_6=1.07555\\
&0.15184l_1+0.18071l_2+0.09918l_3+0.13073l_4-\\
&\qquad 0.01020l_5+1.05154l_6=0.55082
\end{aligned}$$

TABLE 7.—*Percentage of wrong classifications using the function*

$$Z=X_1+X_2+X_3+X_4+X_5+X_6$$

River	Year	147–148	148–149
		Percent	*Percent*
Connecticut	1945	33	22
Hudson	1939	12	22
Hudson	1940	11	19

[2] Only fish with complete meristic data were used; first 91 fish in table 13 (appendix).

TABLE 8.—*Coefficients of the discriminant function* $D_i=\Sigma l_i X_i$ *and successive values of* D_i^2

1	2	3	4	5	6	D_i^2	$D_i/2$	$P(D_i/2)$
1.086	--------	--------	--------	--------	--------	0.4505	0.34	63.3
1.028	0.591	--------	--------	--------	--------	0.6985	0.42	66.3
0.856	0.629	1.029	--------	--------	--------	1.3782	0.59	72.2
0.878	0.630	0.952	0.231	--------	--------	1.4218	0.60	72.6
0.785	0.577	0.871	0:234	1.731	--------	3.1632	0.89	81.3
0.694	0.518	0.850	0.200	1.743	0.246	3.2195	0.90	81.6

The solution of these six equations requires the inversion of a 6 × 6 matrix. Rao (1952) presents a method of solving these equations so that successive discriminant functions are obtained. At the first stage of solution, the discriminant function using anterior scutes only is computed while at the second stage the function using anterior scutes and posterior scutes is obtained. The discriminant function using all six characters is obtained at the sixth stage. The solution of these equations is given in table 8. Any particular discriminant function can be obtained by substituting the l_i from this table into the equation:

$$Y_i=\Sigma l_i X_i.$$

The variance of Y_i is D_i^2 and can be obtained at the same time as the coefficients l_i. It can be proved (Rao 1952) that $D_i/2$ is a normal deviate with mean zero and a standard deviation of one. The probability of obtaining a normal deviate equal to $D_i/2$ is identical to the probability of correctly classifying an individual from any one population. Values of D_i^2, $D_i/2$ and the probability of correct classification are also given in table 8. From this table it can be seen that the increase in D_i^2 with the addition of vertebrae is quite small; therefore, the number of vertebrae is not very useful for purposes of discrimination when used with the other five characters. From the estimates of w_{ij} it is apparent that the covariance between vertebrae and other characters is generally large. This correlation may reduce the usefulness of vertebrae for discrimination. Taking an extreme example where the correlation between two characters is one, it would be useless to include more than one of them in a discriminant function. Immediately the question arises as to how the correlation of the characters affects the relative efficiency of the function. This can be answered by a test of significance which tests the hypothesis of no added increase in D_i^2 in going from a discriminant function using the first p characters to one using p plus q. In this case p=5 and p plus q=6. Rao presents this test on page 253.

$$R=\frac{1+\dfrac{N_1N_2}{(N_1+N_2)(N_1+N_2-2)}D_i^2}{1+\dfrac{N_1N_2}{(N_1+N_2)(N_1+N_2-2)}D_s^2}$$

$$=\frac{1+\dfrac{(91)(104)}{(195)(193)}(3.22)}{1+\dfrac{(91)(104)}{(195)(193)}(3.16)}=1.0078$$

$$F=\frac{N_1+N_2-p-q-1}{q}(R-1)=1.46$$

This F [with q and $(N_1+N_2-p-q-1)$ d. f.] is not significant; therefore, the hypothesis of no added information being supplied by vertebral counts can be accepted. It must be remembered that this is true only when use is made of the data from the remaining five characters. Since vertebrae add nothing to the power of discrimination, they will be omitted from further calculations. The fact that vertebral counts can be eliminated from the discriminant function has considerable practical value, because these counts have to be made from x-rays or after careful dissection of the fish. This one count would probably be as costly in terms of time and money as the other five.

The next step is to find the means of the discriminant function for the two populations. This is done by substituting the mean values of the characters for each population into the discriminant function. The discriminant function as taken from table 8 (excluding vertebrae) is:

$$Y=0.785X_1+0.577X_2+0.871X_3+0.234X_4+1.731X_5$$

The mean value of this function for the Hudson River, 1939, is 74.103 and for the Connecticut River, 1945, is 70.940. If this function were to be used to discriminate between the two populations, those fish with a value of Y less than 72.52 would be called Connecticut River fish and those above 72.52 would be classified as Hudson River fish. The error in this classification would be the proportion of Connecticut fish with a Y greater than 72.52 and the proportion of Hudson fish with a Y less than 72.52. The variance of Y is:

$$D^2=l_1d_1+l_2d_2+l_3d_3+l_4d_4+l_5d_5$$
$$D^2=3.163$$

The proportion of Connecticut River fish which lie in the area under the normal curve from $-\infty$ to 72.52 is equal to the probability of a normal deviate of

$$\frac{72.52-70.94}{\sqrt{3.16}}=\frac{1.58}{1.78}=0.89$$

The probability of this normal deviate is 0.81 (table 8); therefore, the error of misclassification for the Connecticut River population is 19 percent. This is also the error of misclassification for the Hudson River population. This function will correctly classify 81 percent or approximately 3 percent more than the simpler function first investigated.

Rao (1952) presents a test of significance to determine if the calculated discriminant function is better than some other assigned function. If the assigned function is:

$$Z=X_1+X_2+X_3+X_4+X_5+X_6$$

then $$D_Z^2=\frac{(\bar{Z}_1-\bar{Z}_2)^2}{V(Z)}$$

where

$$V(Z)=V(X_1)+V(X_2)+V(X_3)+V(X_4)+V(X_5)+ V(X_6)+2\text{ cov }(X_1X_2)+2\text{ cov }(X_1X_3)+ \ldots +2\text{ cov }(X_5X_6).$$

Using values of w_{ij},

$$D_Z^2=\frac{12.9521}{6.0771}=2.131$$

To test if this function is as reliable as the one derived from the data, the following must be calculated:

$$U=\frac{1+N_1N_2D^2/(N_1+N_2)(N_1+N_2-2)}{1+N_1N_2D_Z^2/(N_1+N_2)(N_1+N_2-2)}-1$$

$$=\frac{1+\dfrac{(104)(91)}{(195)(193)}(3.22)}{1+\dfrac{(104)(91)}{(195)(193)}(2.13)}-1=0.169$$

$$F=\frac{U(N_1+N_2-1-p)}{p-1}=\frac{(0.169)(188)}{5}=6.35$$

F is a variance ratio with $(p-1)$ and (N_1+N_2-p-1) degrees of freedom. In the above instance, F has 5 and 188 degrees of freedom. This is a highly significant value indicating that the calculated function is significantly better than the simpler function.

Since the above discriminant function was based on the 1939 Hudson River sample and the 1945 Connecticut River sample, the 1940 Hudson River sample (table 12, appendix) can be used to demonstrate how the function works. Values for the Hudson River sample of 1940 were substituted in the formula:

$$Y=0.785X_1+0.577X_2+0.871X_3+0.234X_4+ 1.731X_5.$$

The resulting distribution of Y is tabulated in table 9. It can be seen that only 16 out of the 105 values are below 72.52, which is very close to the 19 percent expected. The mean Y for this sample is 74.25, which is in close agreement with the value of 74.10 obtained for 1939.

TABLE 9.—*Frequency distribution of the discriminant function $Y=0.785X_1+0.577X_2+0.871X_3+0.234X_4+1.731X_5$ for the 1940 Hudson River sample*

Y	Frequency	Y	Frequency
78.52–79.51	1	73.52–74.51	10
77.52–78.51	3	72.52–73.51	22
76.52–77.51	4	71.52–72.51	11
75.52–76.51	17	70.52–71.51	4
74.52–75.51	23	69.52–70.51	1

There are a number of assumptions upon which the preceding techniques are based. The two populations have to be multivariate normal populations with equal variances and covariances. It is assumed that the samples are large since sample values are substituted for population values when the discriminant function is calculated. There can be only two populations present, and any future individual that is to be assigned to one of these populations must belong to one of them. Of course if a third population is present with characters considerably different from the two original populations, it may be apparent that it represents a third group when the discriminant function is used.

The calculated discriminant function can be used for two different types of situations. In some studies one is interested in individuals (for example, to obtain scale samples) and would like to be certain that the fish chosen are from an assigned population. In other studies, the relative abundance or composition of a mixed population is desired. In this case there is little interest in the individuals.

If we are interested in classifying individuals, it is possible to adjust the classification region to reduce the chance of making errors. Those individuals that fall close to the division line (75.52) are the cause of the largest percentage of misclassifications. If some of these are not classified, the errors can be reduced. This amounts to dividing the sample into three groups: Hudson River shad, Connecticut River shad, and those that could be either with about equal probability. This third group consists of fish which remain unclassified because there is insufficient information upon which to make a positive identification. If only those fish with a Y less than 70.94 are called Connecticut shad and those with a Y greater than 74.10 are called Hudson River shad, the probability of misclassifying a Conneticut shad would be equal to the area under the normal curve from $74.10-70.94=3.16$ to infinity. The corresponding normal deviate is 1.78 and the area above this value is 3.7 percent. Thus by not classifying approximately one-half of the sample, the number of wrong classifications is reduced to 3.7 percent.

The area of indecision could be extended even wider to further reduce the chance of error; however, if this procedure is carried too far, fish from other rivers might introduce a bias that would have to be considered. The assumption was made earlier that only fish from the Hudson and Connecticut Rivers were present in the sample; however, any fish that do not belong to one of these populations will be classified as though they did. Therefore, any appreciable number of fish from other rivers would cause additional errors. From the tagging experiments mentioned previously, it would appear that a very small percentage of shad present off the New Jersey coast do not belong to one of these two populations. If this is of the order of 5 percent, it might have little effect if all of the fish were classified. If a large portion of the sample remains unclassified, the errors introduced by these fish may be more harmful than those due to misclassifying fish from the two populations.

Estimates of the relative abundance of a mixed population can also be obtained. Three methods of accomplishing this will be presented. The most obvious is to use the discriminant function to classify each fish in the sample and then estimate the composition of the population from the composition of the sample. If there are only two populations present, this method may be quite satisfactory, but it does contain a bias. If a fishery is sampled which contains individuals from only one of these rivers, 19 percent of these fish would be classified as coming from the other race and the estimated composition would be 19 and 81 percent. Thus there would be a bias of 19 percent. If the region is modified so that the relative abundance is estimated from the individuals which are more likely to be classified correctly, then this bias will be reduced. By using the region Hudson>74.10>Unclassified>70.94> Connecticut the estimated composition of a sample which contains only Hudson River fish is

$$\frac{50}{50+3.7}=93.5 \text{ percent}$$

for a bias of 6.5 percent. If there are equal numbers of Hudson and Connecticut River fish present in a sample, then the errors of classification would cancel and the bias would be zero. The maximum bias would occur when a sample is composed of fish from only one river.

Another way of removing the bias is to assume that the error of classification in the sample is the same as the error in the discriminant function (i. e., 19 percent). Then the number of fish classified as Hudson River fish consists of $0.19\ N_C$ and $(1-0.19)\ N_H$ or $N_H=0.19N_C+(1-0.19)N_H$ where N_C and N_H are the numbers present in the population. Similarly for those classified as Connecticut River fish the following relation exists:

$$N_C=(1-0.19)N_C+0.19N_H$$

Substituting sample values (N_C and N_H), these two equations can be solved for N_C and N_H which can be used to determine the relative abundance.

A third estimate is obtained by using the following formula (Rao 1952, p. 300)

$$P=\frac{\overline{X}_H-\overline{X}_S}{\overline{X}_H-\overline{X}_C}$$

where $\overline{X}_{H,}$ $\overline{X}_C$ and $\overline{X}_S$ are the averages of the discriminant function for the Hudson River, the Connecticut River and the sample of the mixed population; P is an estimate of the proportion of the sample native to the Connecticut River.

It is not known which of these estimates would be best for the present problem. A few stray fish will have a greater effect on the first estimate than on the third, particularly if the strays are near one end of the distribution of the discriminant function. In the first type of estimate, they would be weighted more heavily because some of the individuals near the midpoints of the two populations would not be classified. In the third estimate, they would all receive the same weight. In any particular problem, perhaps all three of these estimates should be tried and the various estimates compared. If they are not in agreement, the factors causing the differences should be investigated. Plotting the distributions on probability paper may give some clue to the number of strays present in the samples.

DISCUSSION

The basic condition necessary for the demonstration of a distinct population of shad in each river is that the differences between rivers must be large compared to the differences between years. This condition has been met by the data examined in this study: however, some large differences between years have been reported and they are impossible to evaluate completely at this time. Warfel and Olsen (1947) reported average vertebral counts of 57.042 and 56.837 for 1945 and 1946 in the Connecticut River. This difference of 0.2 is significant. Raney and de Sylva (1953) also reported some differences between years for striped bass. They made the following statement about these differences: (p. 506)

> In any one river system such as the Hudson River there may be significant variations from year to year in any of the characters investigated. These fluctuations may be caused by differences in water temperature and perhaps other factors during larval life at the time when fin ray number is determined. The assumption is made that fin ray numbers are genetically fixed within narrow limits and the minor fluctuations which occur from year to year are due to different physical and perhaps chemical conditions at any one locality or differences in time of spawning, will tend to balance out when samples are taken over a period of several years.

From the statistical point of view, it does not matter what causes these differences when a mixed population is to be divided into its components. For example, the characters for the Hudson and Connecticut Rivers can change considerably from year to year, and the New Jersey catch can still be segregated providing samples are obtained from both rivers and a new discriminant function is calculated each year. Of course, it is essential that the populations be different.

From the biological point of view, the cause of these differences is important. If these differences are primarily genetic, the different populations should be considered taxonomically as races or even sub-species. Raney and de Sylva (1953) considered striped bass from the Hudson River and Chesapeake-Delaware area to be different taxonomic races and suggested calling them the Hudson race and Chesapeake-Delaware race. Similarly, future research may prove that there are actually taxonomic races (or sub-species) of shad.

From Rounsefell and Dahlgren's (1932) work on the herring, it appears that temperature may be one of the most important environmental variables to be studied. A rather simple experiment could be set up whereby it would be possible to hatch shad eggs in controlled water temperatures. This should produce a response curve between meristic counts and temperature, if such a relation exists. Such an experiment would be useful in evaluating the differences between years and rivers.

The human errors in making meristic counts should also be investigated. These certainly contribute to the total variation; therefore, the magnitude of such errors should be known. There are no doubt times when a certain amount of judgment must be used in deciding if a given ray actually should be included in a count. Similarly, gross errors of definition can be made in the counts. These various errors cannot be evaluated at this time, but any future work should certainly include a study of this part of the problem.

Future work with meristic counts will naturally require a great amount of statistical analysis. It is essential, therefore, that the surveys be planned in such a manner that a maximum amount of information can be obtained from them. Of the 1,800 fish collected from 1938 to 1945, only one-third of them could be utilized in a two-way analysis of variance. The surveys should include year classes, sampling dates within a year, different types of gear and different locations within a river. It would also be worthwhile to study the relation between juveniles and the corresponding year class when it enters the fishery as adults. Probably many of the answers which can be obtained from meristic counts will lead to a better understanding of the biology of the various shad populations.

SUMMARY

It is a commonly accepted theory that shad from the different rivers on the Atlantic coast return to the same river to spawn when they reach sexual maturity. Tagging experiments have offered considerable evidence to support this theory. No shad tagged on the spawning ground of one river system has ever been recaptured on the spawning ground of another river system.

If a group of fish return to the same spawning ground year after year with little mixing from other populations, it would be expected that the fish within a river would be more like one another than like the fish from other. rivers. Thus, if differences in some characteristics could be found between rivers, and, if these differences were large compared to the differences between years, the conditions necessary to support a "racial" theory would be present.

Because of the selectivity of the fishing gear used to obtain the samples of shad, it could not be assumed that the samples were random. This selectivity occurred in the size of the fish. The various characters under consideration in this paper were tested for a correlation with length; when no consistent correlations could be found, the samples were considered "representative," even though they were not random.

Analyses of variance of the various characters provided evidence that there were differences between fish from other rivers and, if differences were present between years, they were of a much smaller magnitude than the differences between fish from other rivers. This contributed additional evidence to support the theory of a separate population of shad in each major river.

A large commercial shad fishery exists along the coast of New Jersey, New York Bay, and Long Island. The fishermen in these three areas catch shad that are migrating to the Hudson or Connecticut Rivers. This ocean catch in some years is approximately one-third the size of the river catches and, therefore, should be included in any management plan for the two rivers. To establish a management plan which would include the ocean fisheries would require estimates of the composition of the catches made at these various locations. In the past, this would have had to be done by tagging experiments.

A discriminant function has been constructed in this study which will classify correctly about 81 percent of a mixed population of Hudson and Connecticut River shad. This function was constructed from data obtained from the Hudson River in 1939 and the Connecticut River in 1945. Data from a sample of Hudson River shad obtained in 1940 were substituted into this discriminant function. Out of the 105 fish, 16 were classified incorrectly; this is in good agreement with the theoretical 19 percent misclassifications.

Most of the individuals that are misclassified fall close to the midpoint between the two populations. It is possible to reduce the number of these mistakes by refusing to make a decision on the individuals that lie close to the dividing line between the two populations. This is equivalent to classifying the individuals into three parts: Hudson River, Connecticut River, and a third part for which no decision can be reached. Without using this procedure the chance of misclassifying an individual is 19 percent. By refusing to classify 50 percent of the sample, it is possible to reduce this error to 3.7 percent.

Several methods of estimating the relative abundance or composition of a mixed population are presented. These techniques could be used if one is interested in the population composition of a mixed sample rather than the identification of a particular individual.

LITERATURE CITED

Anderson, R. L., and T. A. Bancroft.
1952. Statistical Theory in Research. McGraw-Hill Publishing Co., New York.

Fredin, Reynold A.
1954. Causes of fluctuations in abundance of Connecticut River shad. U. S. Fish and Wildlife Service, Fish. Bull., vol. 54, pp. 247–259.

Rao, C. Radhakrishna.
1952. Advanced Statistical Methods in Biometric Research. 390 pp. John Wiley and Sons, New York.

Raney, Edward C., and Donald P. de Sylva.
1953. Racial investigations of the striped bass *Roccus saxatilus* (Walbaum). Jour. of Wildlife Management, vol. 17, No. 4, pp. 495–509.

Rounsefell, George A., and Edwin H. Dahlgren.
1932. Fluctuations in the supply of herring, *Clupea pallasii*, in Prince William Sound, Alaska. Bull. U. S. Bur. Fish., vol. 47, pp. 263–291.

Talbot, G. B.
1954. Factors associated with fluctuations in abundance of Hudson River shad. U. S. Fish and Wildlife Service, Fish. Bull., vol. 56, pp. 373–413.

U. S. Fish and Wildlife Service.
1949. Fishery Statistics of the United States, 1945. Statistical Digest 18. 298 pp.

Warfel, Herbert E., and Yngve H. Olsen.
1947. Vertebral counts and the problem of races in the Atlantic shad. Copeia, 1947, No. 3, pp. 177–183.

APPENDIX

Table 10.—*Frequency distribution of meristic counts used in the analysis of variance*

Location	Year	Number of anterior scutes							Number	Mean
		19	20	21	22	23	24	25		
Florida	1938	----	2	9	24	4	----	1	40	21.850
Do	1939	----	1	4	28	7	----	----	40	22.025
South Carolina	1938	----	1	22	12	5	----	----	40	21.525
Do	1939	1	----	14	22	3	----	----	40	21.650
North Carolina	1938	----	1	8	26	5	----	----	40	21.815
Do	1939	----	----	4	28	8	----	----	40	22.100
		Number of posterior scutes								
		12	13	14	15	16	17			
Florida	1938	----	2	12	24	2	----	----	40	14.650
Do	1939	----	1	11	18	10	----	----	40	14.925
South Carolina	1938	1	----	8	14	14	3	----	40	15.225
Do	1939	----	----	9	16	15	----	----	40	15.150
North Carolina	1938	----	1	3	20	15	1	----	40	15.300
Do	1939	----	----	2	12	23	3	----	40	15.675
		Number of anal rays								
		18	19	20	21	22	23			
Florida	1938	1	2	10	18	7	2	----	40	20.850
Do	1939	----	3	9	17	10	1	----	40	20.925
South Carolina	1938	----	4	12	13	9	2	----	40	20.825
Do	1939	----	1	12	18	8	1	----	40	20.900
North Carolina	1938	----	1	8	12	17	2	----	40	21.275
Do	1939	----	3	7	18	10	2	----	40	21.025
		Number of pectoral rays								
		14	15	16	17	18				
Florida	1938	----	2	23	15	----	----	----	40	16.325
Do	1939	----	2	34	4	----	----	----	40	16.005
South Carolina	1938	----	2	18	20	----	----	----	40	16.450
Do	1939	----	4	21	13	2	----	----	40	16.365
North Carolina	1938	----	13	23	4	----	----	----	40	15.775
Do	1939	1	5	30	4	----	----	----	40	15.925
		Number of vertebrae								
		53	54	55	56	57	58	59		
Florida	1938	----	1	2	11	19	7	----	40	56.725
Do	1939	----	----	2	16	19	3	----	40	56.575
South Carolina	1938	----	3	8	18	11	----	----	40	55.925
Do	1939	1	2	5	14	17	1	----	40	56.175
North Carolina	1938	----	----	1	16	19	4	----	40	56.650
Do	1939	----	----	6	6	25	2	1	40	56.650

Table 11.—*Meristic counts on samples of shad from the Hudson River, 1939*

Anterior scutes	Posterior scutes	Dorsal rays	Anal rays	Pectoral rays	Vertebrae
22	15	18	23	15	58
22	15	18	21	16	58
22	15	17	22	14	57
22	15	18	21	16	57
20	15	17	21	16	56
22	16	18	22	17	57
21	16	19	23	15	56
22	16	18	22	16	58
23	14	18	22	16	58
22	17	18	21	16	58
22	16	18	20	15	58
21	15	18	21	17	57
22	16	18	21	17	58
22	16	18	23	15	58
22	17	17	22	15	58
22	15	18	22	17	57
22	16	19	22	15	58
22	15	18	22	15	56
23	15	18	21	15	58
22	15	17	20	16	55
23	16	18	21	16	57
22	16	18	21	16	58
21	16	14	20	15	57
22	14	18	22	15	58
22	16	19	25	16	58
21	15	19	22	16	57
22	15	18	22	15	57
21	13	18	20	15	54
23	17	18	21	16	58
23	15	19	23	16	57
23	15	18	22	17	59
22	15	18	22	16	57
21	15	18	23	16	57
22	16	17	20	16	57
22	15	17	21	16	57
21	15	18	20	16	57
22	14	18	21	15	57
22	16	19	22	17	57
23	16	18	20	16	58
22	14	17	21	16	56
22	16	19	22	16	58
22	15	19	22	15	57
22	14	19	21	16	58
23	16	18	21	16	58
22	16	18	22	16	57
22	16	17	22	16	58
22	14	17	22	15	57
21	17	18	21	17	58
22	18	18	21	15	57
22	15	18	22	15	57
22	15	18	22	16	56
22	15	18	19	16	59
22	15	17	21	17	56
22	16	17	22	15	57
22	16	17	21	16	57
22	15	17	21	16	57
23	15	18	22	16	57
22	15	17	22	16	58
22	15	17	21	16	57
21	15	17	22	15	58
22	14	17	21	15	57
21	14	18	21	15	57
22	15	17	21	16	57
23	15	18	22	16	57
22	15	18	22	15	59
22	16	19	23	16	57

TABLE 11.—*Meristic counts on samples of shad from the Hudson River, 1939*—Continued

Anterior scutes	Posterior scutes	Dorsal rays	Anal rays	Pectoral rays	Vertebrae
22	16	18	22	17	57
22	14	18	20	16	57
22	15	18	21	16	58
22	15	19	22	15	57
22	15	17	21	16	57
23	15	19	22	16	57
23	16	18	22	17	58
22	15	17	23	15	58
22	16	18	22	16	57
21	16	17	21	15	57
22	17	19	22	16	58
22	16	19	20	15	57
22	15	18	22	17	56
21	14	19	22	16	56
22	15	19	22	16	57
23	14	19	21	16	57
22	15	20	23	16	58
23	15	18	22	16	56
22	16	18	22	15	59
22	16	18	22	16	57
21	16	18	22	16	57
21	15	18	22	15	57
22	14	19	23	15	57
22	15	18	22	15	57
22	15	18	21	16	57
22	15	18	22	15	57
23	15	18	23	17	58
21	16	18	22	14	58
22	15	17	21	17	58
22	16	18	22	16	57
22	16	17	21	16	57
23	15	18	20	16	57
23	15	18	21	16	58
23	16	18	22	16	57
23	14	18	22	15	58
23	16	17	23	16	58
22	15	18	23	16	59
22	16	19	22	16	58

TABLE 12.—*Meristic counts on samples of shad from the Hudson River, 1940*

Anterior scutes	Posterior scutes	Dorsal rays	Anal rays	Pectoral rays	Vertebrae
20	15	19	21	16	54
22	16	19	22	17	58
22	16	18	20	17	57
22	16	18	22	16	58
23	15	18	21	16	57
21	16	17	22	15	58
21	15	17	21	15	57
23	16	18	21	16	57
22	16	18	23	16	58
22	15	17	22	15	57
22	15	18	21	16	58
22	16	18	21	17	57
23	16	17	22	16	57
23	14	18	21	15	56
21	15	18	21	16	57
22	15	18	22	15	57
23	15	17	22	15	58
22	16	17	21	16	58
22	16	18	22	15	58
22	15	18	21	16	57
22	16	18	21	16	58
22	16	19	22	16	57
22	17	17	22	15	58
22	15	18	21	15	58
21	17	18	21	16	56
22	15	18	22	15	57
23	16	18	22	16	57
23	15	18	23	16	57

TABLE 12.—*Meristic counts on samples of shad from the Hudson River, 1940*—Continued

Anterior scutes	Posterior scutes	Dorsal rays	Anal rays	Pectoral rays	Vertebrae
22	15	19	23	16	58
20	15	17	22	15	55
22	16	18	21	15	57
22	16	18	22	15	58
23	16	20	24	16	59
22	17	19	20	17	58
22	15	19	24	15	59
22	16	18	22	16	56
21	16	17	21	15	57
22	16	18	21	16	58
22	15	18	21	16	57
22	16	17	21	15	57
21	15	17	20	16	56
22	17	18	22	15	57
23	16	19	22	16	58
22	15	17	22	16	57
22	16	18	23	15	58
23	15	17	19	16	57
22	15	19	21	15	57
22	17	19	21	16	58
22	16	19	21	17	57
22	16	18	21	16	58
22	16	18	22	16	58
23	14	19	22	16	55
22	16	18	21	16	58
22	17	18	21	16	57
22	16	17	20	16	57
22	17	17	20	15	58
23	16	18	22	15	58
21	15	17	22	16	58
23	16	20	23	17	58
22	16	18	21	16	57
23	15	18	21	15	57
22	16	19	21	16	57
22	14	16	20	16	58
22	16	17	21	16	58
22	16	18	22	16	56
21	16	18	21	15	58
24	16	17	20	15	59
23	16	18	22	16	58
23	15	17	21	16	58
24	16	18	21	16	58
22	16	18	22	15	57
22	16	18	22	16	57
22	15	19	23	16	58
22	15	17	20	15	57
22	16	17	23	17	58
22	15	17	23	16	58
22	16	19	21	16	59
21	15	17	23	16	56
22	16	18	20	15	57
21	15	18	22	16	56
22	15	18	20	16	58
22	16	17	22	15	59
22	16	18	22	16	57
23	15	18	21	15	58
22	15	18	22	16	57
22	16	17	22	17	57
23	15	17	20	15	56
22	16	19	20	16	58
22	17	19	22	15	58
22	15	18	22	14	57
22	15	17	20	16	56
23	15	18	22	16	57
22	15	17	23	15	57
22	16	18	21	15	57
22	14	19	22	16	57
23	16	19	23	16	58
21	15	18	23	16	57
22	16	18	22	17	58
22	16	18	22	15	59
22	16	19	24	16	60
22	17	18	22	15	57
22	15	18	21	15	55
22	17	18	22	16	60
22	17	18	20	16	58
22	17	18	22	16	58

TABLE 13.—*Meristic counts on samples of shad from the Connecticut River, 1945*

[The last 10 samples in table are incomplete and therefore were not used in the discriminant function analysis]

Anterior scutes	Posterior scutes	Dorsal rays	Anal rays	Pectoral rays	Vertebrae
21	15	18	21	14	57
22	14	16	20	15	55
21	14	17	21	13	57
22	15	17	21	15	57
22	15	17	20	14	56
21	14	16	22	14	57
22	14	16	21	13	57
21	14	16	20	13	55
22	14	17	21	16	55
22	13	18	23	14	56
22	16	18	21	15	59
22	14	18	22	15	57
22	15	17	20	15	56
20	14	18	22	15	55
22	14	18	23	15	58
21	14	17	22	15	57
22	15	16	21	15	57
21	14	16	20	15	57
21	16	18	19	14	58
23	17	17	21	13	60
22	15	18	20	14	55
23	14	17	22	14	57
21	15	17	22	16	55
21	15	18	22	14	57
22	15	18	21	15	58
23	15	16	19	16	57
21	14	18	21	14	58
22	15	18	21	15	59
22	14	18	22	15	55
21	14	16	22	15	57
22	15	17	21	16	58
21	16	18	21	14	59
21	14	17	21	14	56
23	15	17	21	15	57
21	14	18	22	14	55
22	16	18	20	15	57
21	15	19	21	15	55
22	14	19	21	14	58
20	16	17	21	16	54
21	15	17	21	16	55
22	16	18	20	14	57
22	14	18	21	14	59
22	15	18	22	16	56
22	15	17	20	14	57
21	15	16	21	15	57
21	14	18	21	14	56
21	16	17	22	15	57
21	16	17	22	16	59
22	15	18	21	15	58
21	15	17	21	14	56
22	15	17	21	14	55
22	15	18	21	16	58
22	15	16	22	15	57
21	15	17	22	15	58
22	14	19	22	16	57
21	13	17	21	15	55
21	15	17	21	15	57
22	16	17	21	14	55
21	14	17	20	16	56
22	14	17	20	14	57
22	16	17	21	13	57
21	16	17	23	14	56
22	14	18	21	16	57
21	14	18	23	16	59
22	16	17	23	14	57
22	15	17	22	16	58
21	15	16	20	15	55
22	15	17	22	15	58
21	13	17	23	15	59
21	15	17	21	15	55
22	16	18	23	15	56
22	14	18	21	15	56
22	16	17	20	14	57
21	15	17	21	15	56
23	16	18	21	14	57
22	15	15	19	14	56
22	14	17	20	15	56
22	15	15	20	15	57
22	16	18	23	14	56
22	16	17	23	13	56
22	16	17	21	15	56
21	14	16	23	14	57
22	14	19	21	13	57
21	16	17	21	15	58
22	14	18	21	15	56
21	16	16	21	15	57
21	15	16	22	14	57
22	14	17	22	15	57
21	16	17	21	15	57
21	15	17	22	16	55
22	16	18	22	16	58
22	15	----------	21	15	57
21	15	----------	20	----------	57
21	15	----------	19	14	56
22	15	----------	21	14	57
21	16	----------	----------	13	56
22	15	----------	23	15	57
22	15	----------	----------	14	56
20	16	18	----------	15	57
22	15	17	----------	15	55
22	15	17	----------	14	55