FRI-UW-9902
March 1999

# USE OF STATISTICAL BOOTSTRAPPING FOR SAMPLE SIZE DETERMINATION TO ESTIMATE LENGTH-FREQUENCY DISTRIBUTIONS FOR PACIFIC ALBACORE TUNA (*THUNNUS ALALUNGA*)

M. GOMEZ-BUCKLEY, L. CONQUEST, S. ZITZER, AND B. MILLER

**FINAL REPORT**

**TO**

**NATIONAL MARINE FISHERIES SERVICES**

## ACKNOWLEDGMENTS

We thank Drs. Norm Bartoo and Al Coan at the National Marine Fisheries Service Southwest Center for suggesting the project, providing the data, and funding the project through the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement No. NA67RJ0155. We also sincerely thank Marcus Duke, editor, for editing and formatting this report. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies.

## KEY WORDS

albacore tuna fishery, length-frequency distribution, size determination, *Thunnus alalunga*

## EXECUTIVE SUMMARY

This study used length frequency data collected by the National Marine Fisheries Service from longline fishery vessels from 1962 to 1995. Length-frequency data were collected as tuna were unloaded from longline vessels in American Samoa for the cannery industry. The most abundant species caught in these fisheries was albacore tuna (*Thunnus alalunga*). The main objective of this project was to determine, for the involved countries, the appropriate number of length-frequency samples to collect from landings to obtain representative length-frequency distributions, which would make the data collection process more cost effective for future observer programs.

The data from landings per vessel trip were stratified by country, year, and quarter of the year. Most of the tuna were caught south of the equator, and the samples from north and south of the equator were analyzed separately. A model II ANOVA was used to determine the total variation due to trips. Results showed no significant difference in the length-frequency distributions among the different trips within each stratum. Analysis of histograms revealed the highly varying nature of the length-frequency data, suggesting the use of distribution-free (nonparametric) techniques. Cumulative distribution functions (cdf) were computed for stratified data, and a bootstrapping model based on statistical resampling was used to investigate maximum differences between original stratified distributions and random samples of sizes ranging from 25 to 200. The objective of this procedure was to determine the rate at which the maximum differences between the random sample cdf and the original stratified cdf were reduced. These maximum differences were assessed over the range of 0.025–0.200. For example, preliminary results indicated that 99% of the time, the maximum difference would exceed 0.05 when n = 25, but this percentage is reduced to 58% of the time when n = 200. The applicability of this bootstrapping procedure to a logbook set of data for catch and effort is also discussed in this report.

# Use of Statistical Bootstrapping for Sample Size Determination to Estimate Length-Frequency Distributions for Pacific Albacore Tuna (*Thunnus alalunga*)

M. GOMEZ-BUCKLEY, L. CONQUEST, S. ZITZER[1], AND B. MILLER

## INTRODUCTION

This study used a large sample of length-frequency data collected by the National Marine Fisheries Service from vessels operating in the longline fisheries from 1962 to 1995. Three countries were involved in these fisheries during this period. A code number was used to identify these countries. Length-frequency data were collected as tuna were unloaded from longline vessels in American Samoa for the cannery industry. The most abundant species caught in these fisheries was albacore tuna (*Thunnus alalunga*). The fisheries occurred both north and south of the equator, but 90% of the landings were from south of the equator.

The main objective of this project was to make the process of collecting data more efficient and cost-effective for future observer programs. To attain this objective, the main question was to determine, for the three countries involved in the longline fishery, the appropriate number of length-frequency samples to collect to obtain representative length-frequency distributions for albacore tuna. This was a two-fold question since it was necessary to investigate (1) how many fish lengths to record from each vessel and (2) how many vessels would need to be sampled. This is the first study in which a model was designed to determine appropriate sample size for this type of fisheries.

## MATERIALS AND METHODS

In seeking to statistically define the concept of "adequate representation" of a sample of length-frequency data, we realized that this went beyond the idea of single parameters such as means and variances. For a single sample, we wanted adequate representation across an entire length-frequency distribution. This led us to investigate the behavior of the cumulative frequency distribution (cdf) from a sample of data as compared with the cumulative frequency distribution for a vessel nation-year-quarter. Taking the latter as a defined population against which samples of various sizes would be compared, this led us to consider the approach of the one-sample Kolmogorov-Smirnov test, which looks at the maximum distance between the cdf from the sample and the cdf of the population against which the sample is being tested (Daniel 1990).

The Kolmogorov-Smirnov (K-S) test is designed for use with continuous distributions. For discrete distributions, relying upon tabled critical values for the K-S one-sample test results in a conservative test, that is, one for which the true level of significance is below the stated value. Since a cdf for a given vessel-nation-year-quarter would be a discrete one (although in some instances the steps could be rather fine), we decided to investigate the behavior of the maximum distance based upon the underlying discrete distribution itself, rather than relying upon tabled critical values. Thus, the maximum distance we computed was actually the maximum distance between two step functions—one a random sample from the data set for a particular vessel-nation-year-quarter, the other the entire available data set (taken as a "population") for the vessel-nation-year-quarter.

The statistical software used for the analysis of the data was SPSS version 7.0 (Norusis 1994). The computer used was a DTP 200 mHz processor, 32 MB EDO, and 2 GB Hard Drive. The bootstrapping model was run through the Unix system of the main frame computer at the University of Washington. The total number of entries of albacore length samples was 430,741. The location of the fishery was not recorded 8.5% of the times. A program was designed that could recover this missing information from a presumably parallel set of data for catch and effort, but the logbook data did not yield any matches to speak of for the years 1980 through 1995. An average of 50 albacore

[1]University of Washington Computing and Communications, Seattle, Washington.

were sampled per trip, where "a trip" was defined as the period from when a vessel left a port, to when it arrived at a port to unload the catch.

The variables of interest were extracted from the original data files for each year. The data were stratified by vessel nation (vnation 1 through 3), by year (from 1962 through 1995), and by quarter of the year (quarter 1 = January–March, quarter 2 = April–June, quarter 3 = July–September, quarter 4 = October–December). A total of 278 strata were available (Table 1); however, only a portion of these were used in this study. Spatial stratification was not included in the model. T-tests were run to determine if length frequency data was significantly different for samples north and south of the equator, resulting in samples south of the equator being considered separately from the ones north of the equator. To develop an insight for the amount of variability in the large data sets, exploratory analyses of length frequency data from 31% of the total number of strata were conducted using SPSS summaries of frequency statistics and histogram plots.

A model II ANOVA (see Sokal and Rohlf 1969, p. 211) was used to determine the total variation in the length-frequency distributions due to trips. The question under investigation was to determine if there were significant differences among the length-frequency distributions from different trips. For each stratum there were as many trips as the number of different vessels. A program was designed to determine the number of trips per stratum and to set up

TABLE 1. Summary by year of the vessel nation and quarter of the year represented in the length-frequency data set.

| Year | vnation[1] | Quarter[2] | # strata | Year | vnation[1] | Quarter[2] | # strata |
|---|---|---|---|---|---|---|---|
| 1962 | 1 | 1,3,4 | 3 | 1976 | 2 | 1-4 | 4 |
| | 3 | 1,4 | 2 | | 3 | 1-4 | 4 |
| 1963 | 1 | 1-4 | 4 | 1977 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1964 | 1 | 1-4 | 4 | 1978 | 2 | 1-4 | 4 |
| | 2 | 3,4 | 2 | | 3 | 1-4 | 4 |
| | 3 | 1-4 | 4 | 1979 | 2 | 1-4 | 4 |
| 1965 | 1 | 1-4 | 4 | | 3 | 1-4 | 4 |
| | 2 | 1-4 | 4 | 1980 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1966 | 1 | 1-4 | 4 | 1981 | 2 | 1-4 | 2 |
| | 2 | 1-4 | 4 | | 3 | 1,3,4 | 3 |
| | 3 | 1-4 | 4 | 1982 | 2 | 1-4 | 2 |
| 1967 | 1 | 1-4 | 4 | | 3 | 2,3,4 | 3 |
| | 2 | 1-4 | 4 | 1983 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1968 | 1 | 1-4 | 4 | 1984 | 2 | 1-4 | 4 |
| | 2 | 1-4 | 4 | | 3 | 1-4 | 4 |
| | 3 | 1-4 | 4 | 1985 | 2 | 1-4 | 4 |
| 1969 | 1 | 1-4 | 4 | | 3 | 1-4 | 4 |
| | 2 | 1-4 | 4 | 1986 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1970 | 1 | 1,2,3 | 3 | 1987 | 1 | 4 | 1 |
| | 2 | 1-4 | 4 | | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1971 | 1 | 1-4 | 4 | 1988 | 2 | 1-4 | 4 |
| | 2 | 1-4 | 4 | | 3 | 1-4 | 4 |
| | 3 | 1-4 | 4 | 1989 | 2 | 1-4 | 4 |
| 1972 | 1 | 1 | 1 | | 3 | 1-4 | 4 |
| | 2 | 1-4 | 4 | 1990 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1973 | 2 | 1-4 | 4 | 1991 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1974 | 2 | 1-4 | 4 | 1992 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | | 3 | 1-4 | 4 |
| 1975 | 2 | 1-4 | 4 | 1993 | 2 | 1-4 | 4 |
| | 3 | 1-4 | 4 | 1994 | 2 | 1-4 | 4 |
| | | | | 1995 | 2 | 1-4 | 2 |
| | | | | Total strata | | | 278 |

[1]vnation = vessel nation. Number indicates nation with data.

[2]Quarter 1 = January–March; 2 = April–June; 3 = July–September; 4 = October–December. Numbers indicate quarters for which there are data.

files that would enable data plots and statistical analyses by individual trips.

A bootstrapping model (Efron and Tibshirani 1993) was designed to compute statistics from which the appropriate sample sizes could be established (see Appendix I). The model computed cdf's for each individual stratum (i.e., each year, vessel nation, and quarter of the year). Then the model took random samples (with replacement) from the original cdf (the entire distribution), extracted the maximum distance ($d_i$) between the two cdf's (original and random sample), and repeated this process 1,000 times. The bootstrapping model computed the descriptive statistics, as well as histograms of the maximum distances, in order for us to study their behavior.

The random samples taken from the original cdf were of the following sample sizes: n = 25, 50, 100, 150, and 200. These sample sizes represented the number of fish that would be sampled per vessel nation, and quarter, to record length measurements. The maximum distances were assessed over the range of 0.025–0.200. The objective of this procedure was to determine the rate at which the maximum distances between the random sample cdf and the original cdf were reduced. Tables that contained the proportion of the time that the maximum distances exceeded a stated distance, for a given sample size (an average of 10 trials for a given sample size), were produced for each stratum. Random number seeds were used for each of the 10 sets of each sample size (a total of 50 per run or stratum). The seed would initialize a random number generator; the seed needed to be greater than 1 and less than 2,000,000.

## RESULTS

The number of albacore length samples ranged from 4,682 north of the equator to 389,376 south of the equator; there were 36,683 samples from unknown locations (Fig. 1). From all years investigated (1980–95) for possible recovery of data from unknown locations, the program was able to recover a total of 81 entries from the catch and effort dataset. The catches from south of the equator accounted for 90% of the length-frequency data. The results from the t-tests run for years 1988 and 1989 showed that length samples from north and south of the equator were significantly different (a = 0.05).

Histograms were plotted for 88 randomly chosen strata (32% of the total number of strata). Examples of the variability in the length-frequency distributions are demonstrated by the data for stratum vnation 2, 1980, quarter 1, and stratum vnation 2, 1980, quarter 2 (Fig. 2a and b, respectively). The distribution in Figure 2a is right skewed

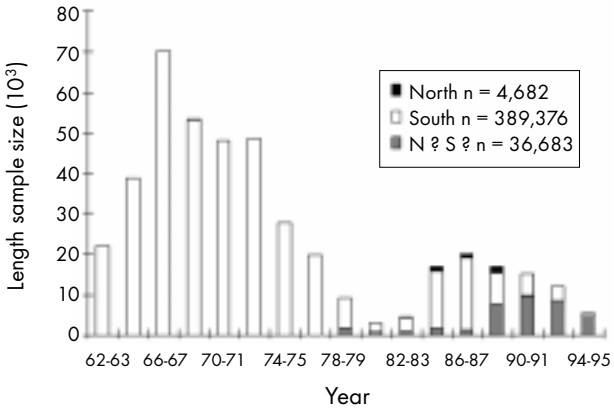


FIGURE 1. Sample size for albacore lengths north and south of the equator (1962–95).

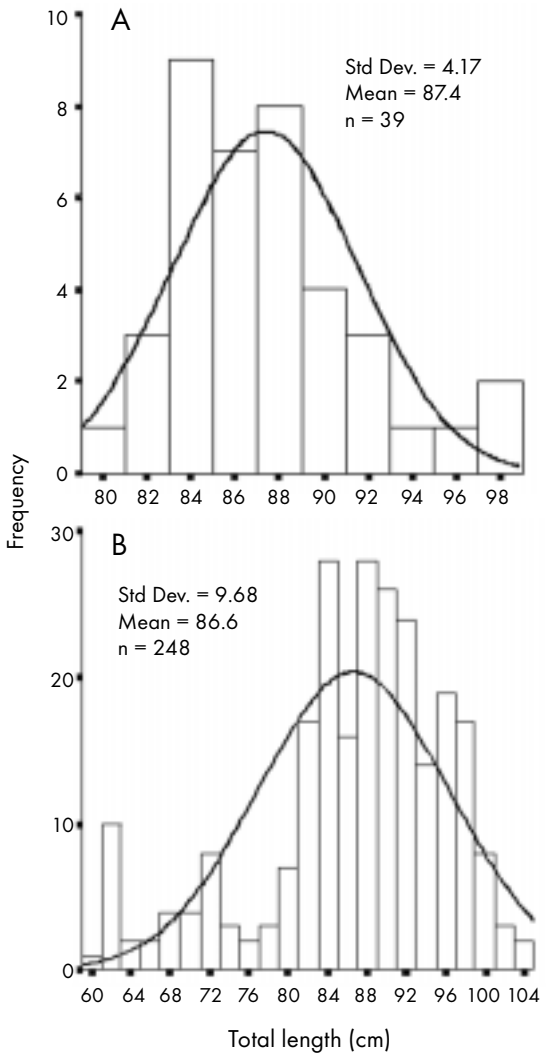


FIGURE 2. Histogram of length-frequency distributions: (A) nation 2, year 1980, and quarter 1; (B) nation 2, year 1980, quarter 2. Smoothed curve is a fitted normal distribution.

while the one in Figure 2b is left skewed. Sample sizes also differed considerably. The remainder of the exploratory analyses of length frequencies, by year, for 1962 through 1995 is given in Appendix II. Figures 3a and b are the cumulative distribution functions for Figures 2a and b, respectively. Again, they show a marked difference in the shape of the distributions for each stratum. Analysis of histograms and descriptive statistics revealed the highly varying nature of the length-frequency data; this suggested the use of distribution-free (nonparametric) techniques.

The model II ANOVA showed no significant difference in the length frequencies among the different trips within each stratum. The percentages of the variation due to trips ranged from 2.3% to 19% (Table 2).

For each stratum, the bootstrapping program generated 1,000 maximum distances; Figure 4 illustrates how a single maximum distance was extracted for each random sample's cdf. The flow chart in Figure 5 shows, in a simplified form, the steps to get to the maximum distances. An example is given for 1,000 bootstrap runs of a random sample of n = 100 per run for a generic stratum (Fig. 6) where the amount of the skewness in the maximum distance distributions decreased as the sample size increased. An example is also given for random samples of n = 25, 100, and 200 (Fig. 7).

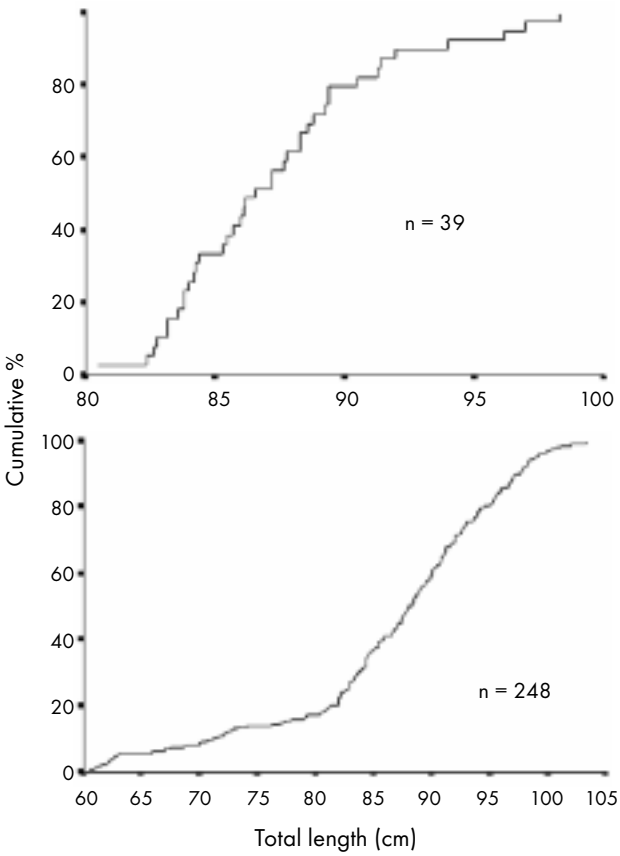


FIGURE 3. Cumulative distributions functions: (A) nation 2, year 1980, quarter 1; (b) nation 2, year 1980 quarter 2.

TABLE 2. Percentages of the variation due to trips for several strata from the model II ANOVA.

| Stratum | % var. comp. trips |
|---|---|
| vn1y62q1 | 13.8 |
| vn1y62q3 | 19 |
| vn1y62q4 | 15 |
| vn3y62q1 | 7 |
| vn3y62q4 | 2.3 |

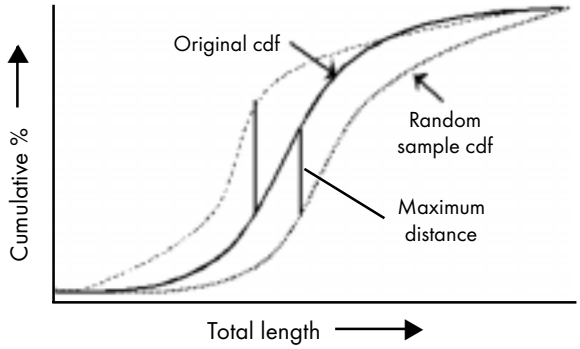


FIGURE 4. Maximum distance between original and random sample cumulative distribution functions.

The results of the bootstrap runs for year 1980, vnation 2, quarters 1 and 2 (Tables 3a and b) showed strong similarity (within 0.03 difference) for most of the values for the proportion of the times that maximum distances exceeded a stated distance for a given sample size. The largest difference (0.07) is for stated distance 0.05 and random sample n = 150. For year 1990, vnation 2, quarters 1 through 3 (Table 3c, d and e), the values are also very similar, with again a maximum difference of 0.07 for stated distance 0.05 and random sample n = 200.

## DISCUSSION

From the results of the ANOVA II model, it appears that it would not be necessary to sample more than one trip for each quarter of the year. The sample size selected from the bootstrapping model would be the total number of fish necessary to obtain a representative length-frequency distribution for a specific nation, year, and quarter.

The bootstrapping model was run successfully five times. The inspection of the histograms of the maximum distances showed that they were right skewed. The amount of the skewness depended upon the sample size; that is, as n got larger the skewness decreased. In order to make any recommendations for the appropriate sample size for length fre-

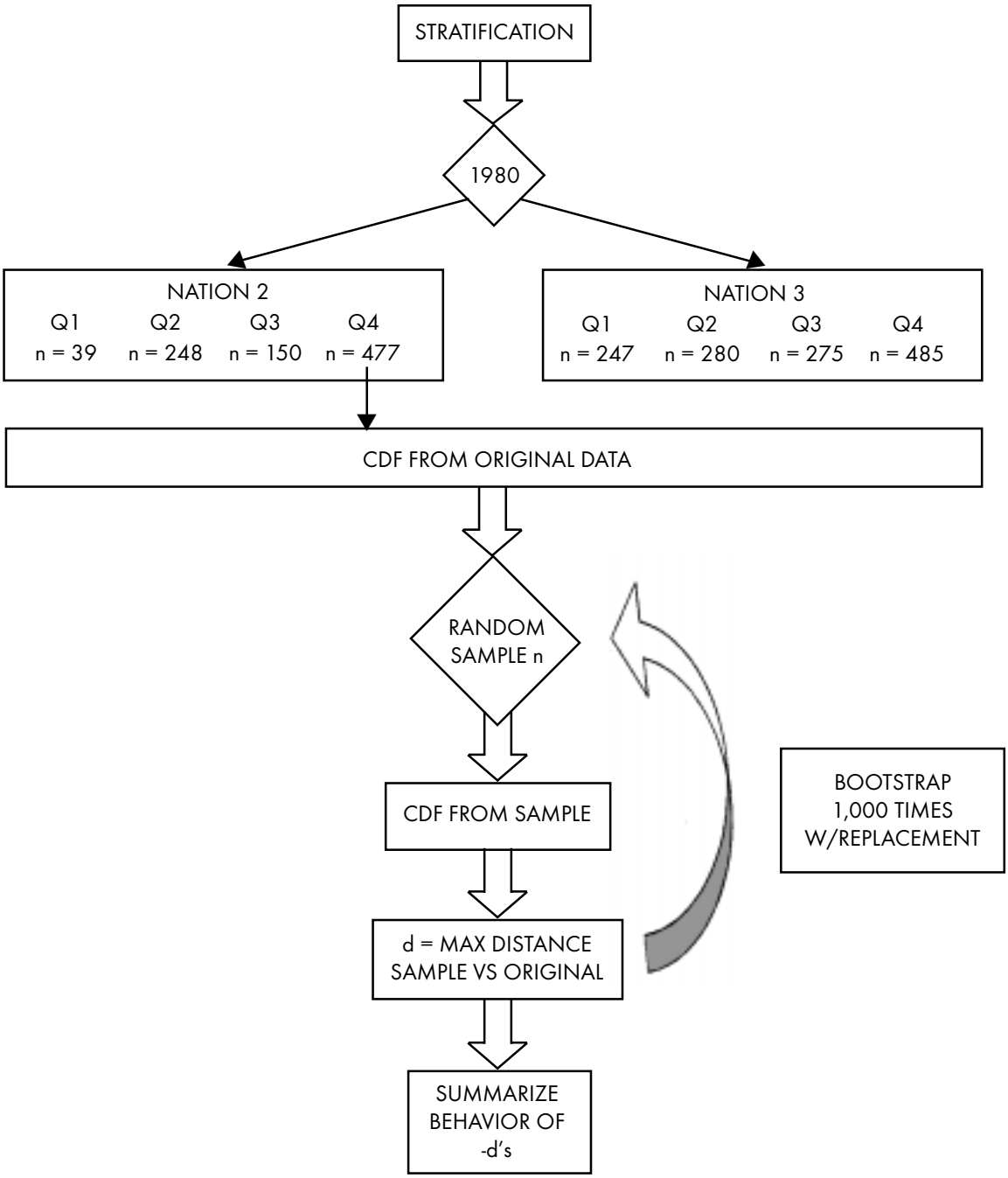


FIGURE 5. Steps of the process to extract the 1,000 maximum distances from a single stratum.

quency distributions of albacore tuna, we would have preferred to run the model many more times (the total number of strata was 278). Preliminary results from the five runs of the model (Tables 2a–e) indicated that 98% of the time the maximum difference between the original cumulative distribution function and the random sample cumulative distribution function will exceed 0.05 when n = 50, but reduces to 53% of the time when n = 200. The stated distance of 0.15 was never exceeded for samples sizes of n = 150.

The process of designing the model that determined appropriate sample size for length frequency distributions of albacore tuna was complex. It was time-consuming to enter the values for each stratum in the bootstrapping model, and once these were entered, it took approximately 2 hours to run the model in the Unix system. To make the whole process faster, a subroutine could be introduced into the model that would generate random numbers for the seeds that would otherwise have to be entered manually

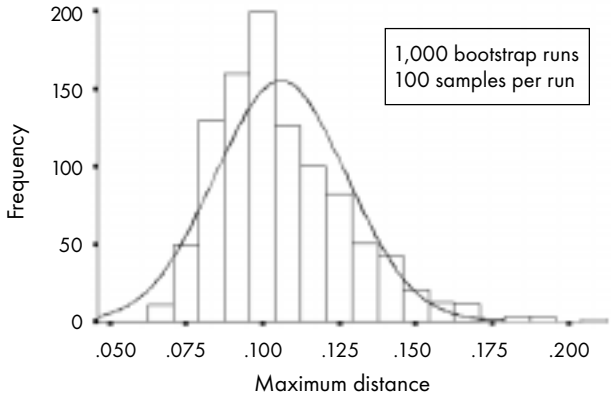


FIGURE 6. Frequency distribution of maximum distances ($d_i$) for a specific stratum and random sample. Smoothed curve is a fitted normal distribution.

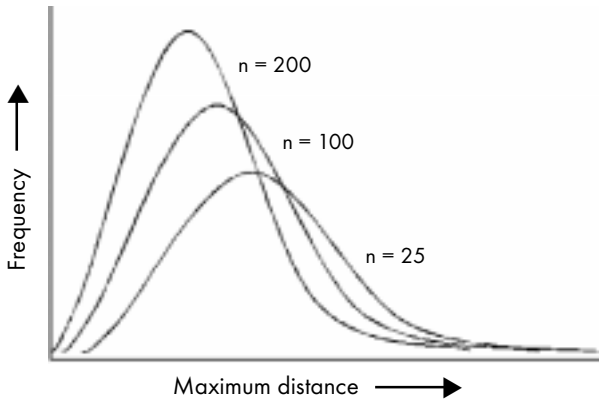


FIGURE 7. Diagram of maximum distance distributions for random samples of 25, 100, and 200.

into the model one by one (a total of 50 for each stratum). Instead of conducting 1,000 bootstrap runs of the model 10 times for each random sample, the model could be run once for each random sample. This was based on the preliminary analyses of the descriptive statistics and histograms of the 10 sets of 1,000 bootstraps for each random sample, which appeared to have similar behavior. This would increase the output of the model considerably, since the number of seeds would significantly be reduced.

One of the main problems when dealing with this type of data was the way the logbooks were recorded. The fact that the fish lengths were taken at the time of unloading the catches in port, while the locations were taken from the vessel logbooks, created the possibility for not having accurate data for the times and exact locations of the harvests.

One should also be able to apply the bootstrap simulation procedure discussed in this report to catch or effort (or catch-per-unit-effort) data from logbooks. To assess the feasibility of this approach, we examined histograms for albacore catch and effort (number of hooks used) for eight combinations of vessel nation, year, and quarter. For albacore catch, as expected, varying degrees of right skewness were observed. The histograms for effort ranged from "reasonably symmetric" to left or right skewed. It should be possible to replicate our bootstrap exercise and determine how many entries would be needed to ensure that the maximum distance between the designated "population" cdf and the sample cdf is kept at or below a set value with a high probability. Then, since logbooks would be expected to have a varying number of entries, one would need to examine the frequency distributions for catch and effort over a number of logbooks. (For example, an analysis of variance could be done to assess logbook variability with respect to catch and effort, keeping in mind that vessel nation, year, and quarter are also other sources of variation. Since the concern is regarding the entire distribution of catch and effort and not just average catch and effort, however, an approach involving comparison of entire frequency distributions is probably in order. The particular analysis done would depend somewhat upon what the data showed.) The more homogeneity among logbooks, the fewer that would have to be sampled.

## LITERATURE CITED

Daniel, W. 1990. Applied Nonparametric Statistics. PWS-Kent Publishing Company, Boston. 635 p.

Efron, B. and R.J. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman Hall, New York. 436 p.

Norusis, M.J. 1993. SPSS for Windows: Base System User's Guide, Release 6.0. SPSS Inc. 828 p.

Sokal, R.R. and F.J. Rohlf. 1969. Biometry. W.H. Freeman and Company, San Francisco. 776 p.

TABLE 3. Proportion of the times that maximum distances exceed a stated distance for a given random sample size for strata: (A) 2-80-1, (B) 2-80-2, (C) 2-91-1, (D) 2-91-2, and (E) 2-91-3. The row entries are the stated distances of interest, the column entries are the varying sampling sizes, and the entries inside the tables are the proportion of the times that the maximum distances meet or exceed this stated distances. Each entry on the table was the average from 10 sets of 1000 bootstrap runs.

**A**

| 2-80-1 | Random sample size (n) | | | | |
|---|---|---|---|---|---|
| Max. Dist. | 25 | 50 | 100 | 150 | 200 |
| 0.025 | 1 | 1 | 1 | 1 | 0.99 |
| 0.05 | 1 | 0.98 | 0.85 | 0.67 | 0.53 |
| 0.075 | 0.93 | 0.78 | 0.45 | 0.25 | 0.13 |
| 0.1 | 0.78 | 0.49 | 0.17 | 0.05 | 0.02 |
| 0.125 | 0.58 | 0.25 | 0.05 | 0.01 | 0 |
| 0.15 | 0.38 | 0.11 | 0.01 | 0 | 0 |
| 0.175 | 0.24 | 0.05 | 0 | 0 | 0 |
| 0.2 | 0.14 | 0.02 | 0 | 0 | 0 |

**B**

| 2-80-2 | Random sample size (n) | | | | |
|---|---|---|---|---|---|
| Max. Dist. | 25 | 50 | 100 | 150 | 200 |
| 0.025 | 1 | 1 | 1 | 1 | 0.99 |
| 0.05 | 1 | 0.99 | 0.89 | 0.74 | 0.58 |
| 0.075 | 0.95 | 0.82 | 0.49 | 0.28 | 0.15 |
| 0.1 | 0.81 | 0.53 | 0.19 | 0.07 | 0.02 |
| 0.125 | 0.62 | 0.28 | 0.06 | 0.01 | 0 |
| 0.15 | 0.42 | 0.13 | 0.02 | 0 | 0 |
| 0.175 | 0.26 | 0.06 | 0 | 0 | 0 |
| 0.2 | 0.15 | 0.02 | 0 | 0 | 0 |

**C**

| 2-91-1 | Random sample size (n) | | | | |
|---|---|---|---|---|---|
| Max. Dis. | 25 | 50 | 100 | 150 | 200 |
| 0.025 | | | | | |
| 0.05 | 1 | 0.96 | 0.81 | 0.62 | 0.47 |
| 0.075 | | | | | |
| 0.1 | 1 | 0.45 | 0.15 | 0.04 | 0.01 |
| 0.125 | | | | | |
| 0.15 | 0.94 | 0.1 | 0.01 | 0 | 0 |
| 0.175 | | | | | |
| 0.2 | 0.79 | 0.01 | 0 | 0 | 0 |

**D**

| 2-91-2 | Random sample size (n) | | | | |
|---|---|---|---|---|---|
| Max. Dis. | 25 | 50 | 100 | 150 | 200 |
| 0.025 | | | | | |
| 0.05 | | 0.98 | 0.86 | 0.68 | 0.53 |
| 0.075 | | | | | |
| 0.1 | | 0.47 | 0.17 | 0.06 | 0.02 |
| 0.125 | | | | | |
| 0.15 | | 0.12 | 0.01 | 0 | 0 |
| 0.175 | | | | | |
| 0.2 | | 0.02 | 0 | 0 | 0 |

**E**

| 2-91-3 | Random sample size (n) | | | | |
|---|---|---|---|---|---|
| Max. Dis. | 25 | 50 | 100 | 150 | 200 |
| 0.025 | | | | | |
| 0.05 | | 0.97 | 0.86 | 0.66 | 0.54 |
| 0.075 | | | | | |
| 0.1 | | 0.47 | 0.16 | 0.06 | 0.03 |
| 0.125 | | | | | |
| 0.15 | | 0.12 | 0.01 | 0 | 0 |
| 0.175 | | | | | |
| 0.2 | | 0.01 | 0 | 0 | 0 |