

Sample-Size Requirements for Evaluating Population Size Structure

J. C. VOKOUN

*Department of Fisheries and Wildlife Sciences, University of Missouri,
302 Anheuser-Busch Natural Resources Building,
Columbia, Missouri 65211-7240, USA*

C. F. RABENI*

*U.S. Geological Survey, Missouri Cooperative Fish and Wildlife Research Unit,
University of Missouri, 302 Anheuser-Busch Natural Resources Building,
Columbia, Missouri 65211-7240, USA*

J. S. STANOVICK

*Missouri Department of Conservation, Columbia Research Center,
1110 South College Drive, Columbia, Missouri 65201, USA*

Abstract.—A method with an accompanying computer program is described to estimate the number of individuals needed to construct a sample length-frequency with a given accuracy and precision. First, a reference length-frequency assumed to be accurate for a particular sampling gear and collection strategy was constructed. Bootstrap procedures created length-frequencies with increasing sample size that were randomly chosen from the reference data and then were compared with the reference length-frequency by calculating the mean squared difference. Outputs from two species collected with different gears and an artificial even length-frequency are used to describe the characteristics of the method. The relations between the number of individuals used to construct a length-frequency and the similarity to the reference length-frequency followed a negative exponential distribution and showed the importance of using 300–400 individuals whenever possible.

Length-frequency distributions are used to evaluate fish population characteristics such as age structure (Tesch 1971), reproduction, recruitment, growth, and mortality (Anderson and Neumann 1996). To efficiently use size structure data sets, fisheries biologists must know how many fish need to be collected to build a reliable length-frequency histogram. The use of abbreviated length-frequency distributions to calculate ratios that characterize fish populations such as proportional stock density (Anderson 1978) and relative stock density (Wege and Anderson 1978) has been evaluated, and confidence intervals based on sample size have been reported (Gustafson 1988). Anderson and Neumann (1996) suggest that at least 100 stock-sized fish be used.

However many uses of length-frequency histograms require the information contained in the entire distribution, yet we found no general methodology or recommendations in the literature. In these instances, biologists would benefit from knowing the accuracy and precision of length-frequency histograms constructed at a particular sample size relative to greater sample sizes.

We present a method for determining the number of fish (sample size) required to produce a length-frequency data set that is representative of the sizes of fish available to be collected by a particular gear under particular conditions. The examples used here are each relative to a reference length-frequency delivered by a gear used in a standardized manner. The relationship of that length-frequency to the actual population of fish may or may not be known depending on previous quantification of the size bias and efficiency of the gear used.

Methods

A reference length-frequency (relative and based on percent of total) was created for a species by the use of data from an initial sampling in which an unusually large number of fish were collected using gears and methods identical to those normally employed. New relative length-frequency distributions determined with increasing sample size were created from the reference population by random computer sampling with replacement. The reference length-frequency A was then compared with a sample length-frequency B as follows: for each length interval, the percent difference between A and B was squared, and the mean value among all length intervals was calculated. The re-

* Corresponding author: rabenic@missouri.edu

Received May 8, 2000; accepted November 21, 2000

sultant value is the mean squared difference (MSD) between A and B. The value of MSD indicates the relation between A and B (i.e., the smaller the MSD, the greater the similarity between the two length-frequency histograms). Because MSD describes the similarity of two length-frequency histograms, it can be used to reveal the effect of sample size on length-frequency histogram generation.

The SAS computer program (Appendix 1) was used to randomly sample (with replacement) fish lengths from the reference data at 12 preselected sample sizes (50, 100, 200, 300, . . . 1,300, 1,400, and 1,500) and then to compare the new computer-generated length-frequency with the reference length-frequency by calculating the MSD value. The program can be configured to use any size increment or length desired. Because each time the computer program randomly sampled the reference data and created a length-frequency that was a unique, equally probable event, we repeated the process and created 100 length-frequency distributions (and ultimately 100 MSD values) for each sample size and assembled an empirical distribution. This resampling process is known statistically as bootstrapping (Efron and Gong 1983). The SAS program is configured to output the mean MSD, median MSD, and the first and third quartiles, as well as the 5th and 95th percentiles (these values can be adjusted).

We illustrate this method for two species captured with different gears, channel catfish *Ictalurus punctatus* ($n = 2,761$) collected with hoop nets in the Grand River basin of Missouri and bluegill *Lepomis macrochirus* ($n = 654$) captured by standardized boat electrofishing on Ashland Lake in central Missouri, and with an artificial evenly distributed length-frequency ($n = 3,400$). The length-frequency histograms of these data sets (Figure 1) were constructed with both 10- and 25-mm length intervals.

Results and Discussion

General Relationship

As sample size increased, the mean MSD value decreased, and computer-generated length-frequency distributions became increasingly similar to the reference length-frequency in all data sets (Figure 2). Relations between sample size and MSD followed a negative exponential model and became asymptotic to the x -axis. The first several increases in sample size reduced the MSD much more than subsequent increases. The MSD can be

used as an index of the accuracy of a sample size. Accuracy is defined here as how similar the computer-generated length-frequency distributions are to the reference length-frequency as measured by the MSD. Different length-frequency measurements required different sample sizes to achieve the same level of accuracy (Figure 2).

Effect of the Number of Length Intervals

The evenly distributed artificial length-frequency was analyzed by the use of 17 and 34 length intervals. Resultant mean MSD values from the 17 length-interval (25-mm) bootstrapping runs were approximately half those of the 34 length-interval (10-mm) results, indicating greater accuracy. This general pattern was repeated in the bluegill data set; however, as sample size increased, the effect was less important because values converged within the 90% inclusive data band, i.e., the 5th and 95th percentiles. The channel catfish data set results were different; mean MSD was actually similar or slightly higher for the 25-mm-interval bootstrapping runs. This occurred because of the shape of the length-frequency (see following section). Generally precision (represented by the 90% inclusive data band) increased when more length categories were used, although other characteristics of the length-frequency can interact and mask the effect. However the 90% inclusive data band overlapped to a large extent within the 10- and 25-mm results in all data sets analyzed, indicating that the increase in precision was not substantial.

Effect of Length-Frequency Shape

As the reference length-frequency deviates from an even distribution, accuracy and precision decrease for a given sampling effort. The reason for this is analogous to stacking the deck in a card game. For example, if half of the diamond suit were removed from the deck and replaced with the corresponding hearts, the greater frequency of hearts in the deck would require that more cards have to be sampled (assuming replacement) to represent all cards present. The unequal probabilities that sampled individuals will belong to a particular length category accounted for the results seen in the channel catfish data set. When the data set was analyzed at 10-mm length increments, no length category accounted for more than 13% of the length-frequency distribution. However, at 25-mm increments, the 126–150-mm length category accounted for 24% of the total individuals in the length-frequency, effectively moving the distri-

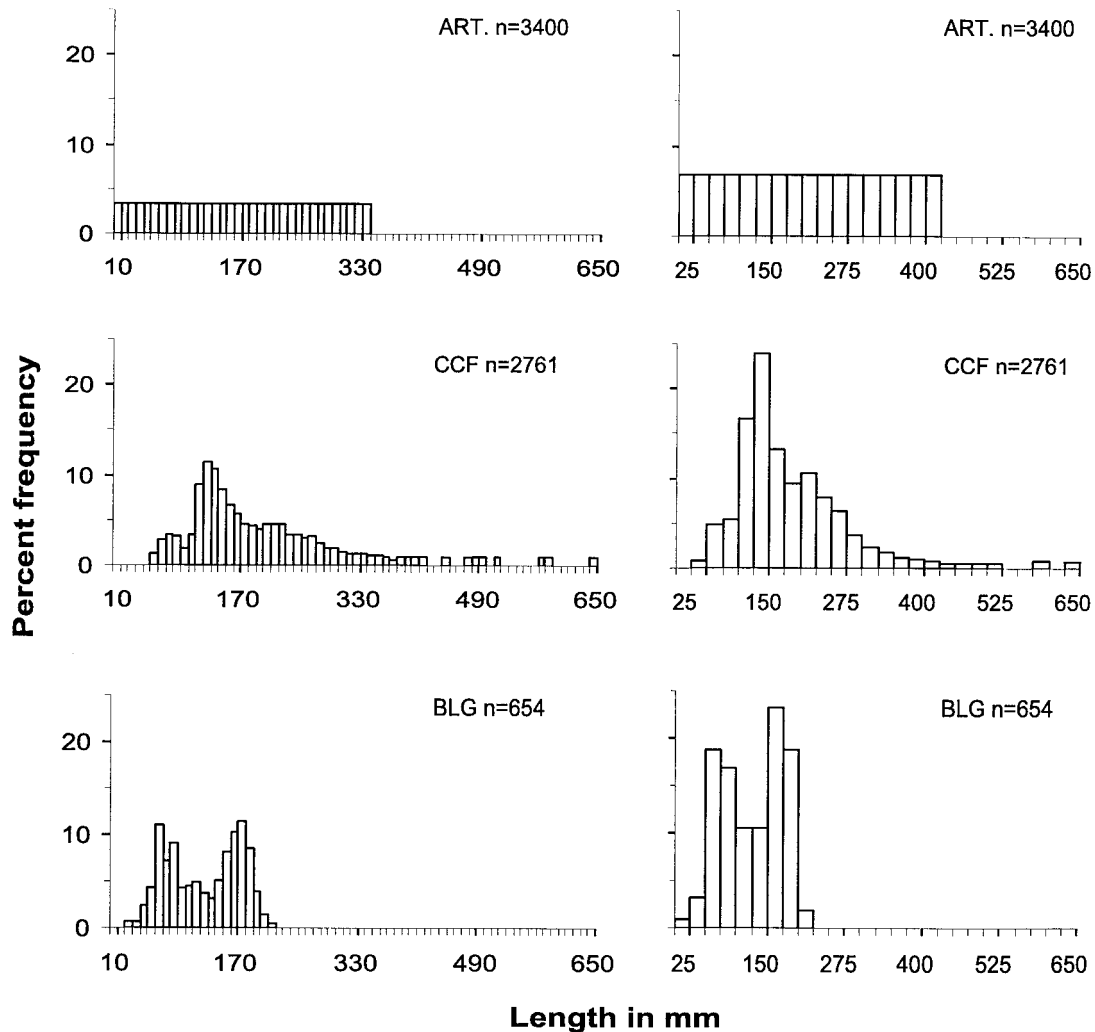


FIGURE 1.—Length-frequency histograms for artificial (ART), channel catfish (CCF), and bluegill (BLG) data sets. Both 10-mm (left column) and 25-mm (right column) length categories are presented.

bution farther away from even. Length categories represented by a small percentage of the total length-frequency distribution were less likely to be accurately represented in samples because of the increased probability that sampled individuals would belong to a length category better represented. This increased disparity in probabilities that an individual sampled would belong to a particular length category resulted in an increased mean MSD. This effect was not seen in the bluegill data set because only 3 (or 33%) of the length categories present were represented by small percentages of the total length-frequency, whereas 11 (or 50%) of the length categories in the channel catfish length-frequency were represented by small

percentages. In length-frequency distributions shaped like that of our channel catfish population (Figure 1), which have many length intervals represented by small percent frequencies, biologists may want to place less emphasis on these intervals by manually combining them into one interval. For example, in the channel catfish population, all fish over 400 mm could be assigned a length of 401 mm, creating an over-400-mm interval that would be represented in the reference length-frequency and sampled as a single interval.

Effect of Maximum Fish Size

Channel catfish returned greater MSD values than the smaller bluegill (Figure 2). This occurs

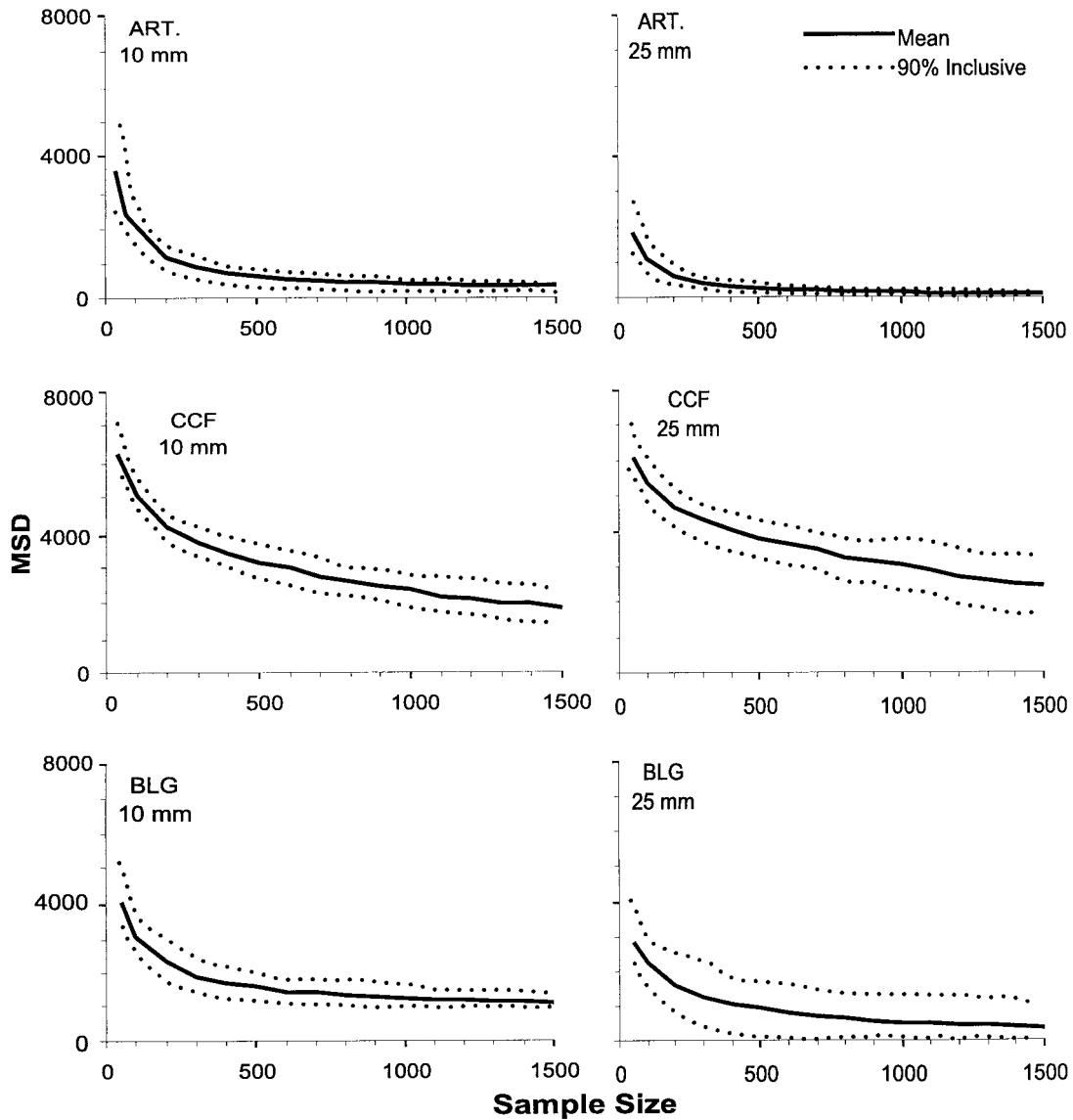


FIGURE 2.—Relation of mean squared difference (MSD) and the 90% inclusive data band to sample size for artificial (ART), channel catfish (CCF), and bluegill (BGL) length-frequency data sets.

for the same reason discussed for length interval number. Length data for a larger, longer-lived species will most likely be divided into more length intervals than the data for a smaller species, even though the smaller species would be divided into smaller (fewer mm) length intervals. A larger species inherently cannot be sampled as accurately as a smaller-bodied one without increased effort. Decreasing the number of length categories (e.g., switching from 10- to 25-mm increments) will reduce the MSD; however, the use of length-

frequency data is for population analysis, and we suggest that length intervals should be based on predetermined meaningful growth increments.

Recommendations

We present a method to determine appropriate sample sizes for length-frequency analysis of particular species of interest, given specific objectives and resources. The data sets presented here suggest sample sizes of 300–400 begin to show a leveling off of the decrease in MSD. We also are aware of

instances in which the collection of a sufficient reference length-frequency is not possible (e.g., evaluating historical data). We believe that in those cases this method, used on similar taxa with similar sampling techniques, will describe the level of variation contained in a small sample and the accuracy relative to a larger sample size and can thus provide a valuable context for management decisions. This method assumes a large or open population within which sampling at the scale described here (hundreds of individuals) would not change the relative availability of length classes. This scenario is more common than not in rivers and streams where sampling is often linear in nature with little resampling of areas but is certainly contextual. Fisheries biologists will need to exercise professional judgment to decide for themselves whether this program simulates their sampling situation. For example, if a manager were interested specifically in a 3-km reach of a wadeable stream, this method is an appropriate simulation for most of the common darters and minnows but may not be an accurate simulation for smallmouth bass *Micropterus dolomieu*. However, if the sampling were increased in scale and conducted throughout a watershed or along a larger reach, this method could adequately simulate smallmouth bass sampling. Coastal estuaries and nearshore marine fisheries also have large populations that are routinely sampled. Lentic systems such as natural lakes and reservoirs constitute other situations that are suitable for simulation by the sampling program presented here, although the smallest lakes and impoundments should be avoided.

Acknowledgments

We gratefully acknowledge C. Gale who developed and wrote a precursor to the SAS program

used in this manuscript; R. Hayward provided the bluegill data set. Major funding was provided by the Missouri Department of Conservation. This is a contribution from the Missouri Cooperative Fish and Wildlife Research Unit (U.S. Geological Survey, Missouri Department of Conservation, University of Missouri, and Wildlife Management Institute cooperating). This manuscript was improved with comments from three anonymous reviewers.

References

- Anderson, R. O. 1978. New approaches to recreational fishery management. Pages 73–78 in G. D. Novinger and J. G. Dillard, editors. New approaches to the management of small impoundments. American Fisheries Society, North Central Division, Special Publication 5, Bethesda, Maryland.
- Anderson, R. O., and R. M. Neumann. 1996. Length, weight, and structural indices. Pages 447–482 in B. R. Murphy and D. W. Willis, editors. Fisheries techniques, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross validation. *The American Statistician* 37:36–48.
- Gustafson, K. A. 1988. Approximating confidence intervals for indices of fish population size structure. *North American Journal of Fisheries Management* 8:139–141.
- Tesch, F. W. 1971. Age and growth. Pages 98–130 in W. E. Ricker, editor. *Methods for assessment of fish production in fresh waters*. Blackwell Scientific Publications, London.
- Wege, G. J., and R. O. Anderson. 1978. Relative weight (Wr): a new index of condition for largemouth bass. Pages 79–91 in G. D. Novinger and J. G. Dillard, editors. *New approaches to the management of small impoundments*. American Fisheries Society, North Central Division, Special Publication 5, Bethesda, Maryland.

Appendix: Computer Program

We have included the SAS program used to examine the data in this manuscript. The program should be run at each of the sample sizes (e.g., 50, 100, 200, . . . 1,500) to generate the mean squared difference values needed to construct a trendline graph as in Figure 2. Text in italics is explanative and indicates where in the program the user needs

to provide information. Personal computer SAS system for Windows users will need to exit the program and begin a new session after each run because the program will continue to append the results, giving a cumulative value rather than the correct value at a given sample size.

The program is as follows:

```
%macro boostr;
  %do j = %to 100;
    data samptbl;
      drop i;

do i = 1 to _; * enter the sample size;
  iobs = int(ranuni(0) *N) + 1;
  set ds0 point=iobs nobs=N;
  output
  end;
stop;

proc freq data=samptbl;
  tables length / noprint out=samptbl;

data tbl;
  merge samptbl poptbl;
  by length;
drop count;
  if percent = . then percent = 0;
  if percent = 0 then diffsq = (100)**2;
  if percent ge stdpct then diffsq = (((stdpct / percent) * 100) - 100)**2;
  if stdpct gt percent then diffsq = (((percent / stdpct) * 100) - 100)**2;
proc means data=tbl noprint;
  var diffsq;
  output out=difsqb mean =msqdif;
proc append base=difsq data=difsqb;

%end;
%mend boostr;

data ds0;
  input L;
  length = round (length, —); * enter your chosen length category size;
infile —; * enter your reference length data;
proc freq data=ds0;
  tables length / noprint out=poptbl;
data poptbl;
  set poptbl;
  rename percent=stdpct;
  drop count;
data dummy;
  set ds0;
  % boostr;
proc sort data=difsq out=difsq;
  by msqdif;
proc print data=difsq;
proc univariate noprint data=difsq;
  var msqdif;
  output out=dsprint mean=mean median=median p5=p5 q1=q1 q3=q3 p95=p95;

proc print data=dsprint;
run;
QUIT;
run;
```