

255
BIB 35
#3

APPENDIX A

GEORGETOWN LAKE FISHING STUDY-1969 WINTER ICE FISHING SEASON
SURVEY DESIGN AND STATISTICAL CONSIDERATIONS

Prepared by Kenneth P. Johnson

INTRODUCTION

Ice fishing will be allowed on Georgetown Lake for the ten week period beginning Sunday, December 21, 1969 and ending Saturday, February 28, 1970. The fishing hours will run from 5 a.m. to 10 p.m. on all days.

It is desired to collect data during this season to provide estimates of the following related to Georgetown Lake:

1. Total fishing pressure for the season (number of hours of effort)
2. Total fish harvest for the season
3. Total number of fisherman trips made during the season
4. Average fish harvest per hour of effort (fish per hour)
5. Average fish harvest per fisherman trip (fish per man)
6. Average length of fisherman trip (hours per man)

The estimates of fish per hour, fish per man, and hours per man will be derived through data gathered from fishermen who have completed their fishing trips. The estimates of the totals must be made by estimating the average number of cars parked on the lake shore and relating this number to

fishermen through an estimate of the average number of fishermen per car. Thus, we need observations on the number of cars parked on the lake shore at random times and interviews with a sample of fishermen.

It is the intent of this paper to describe the following:

1. The plan used to collect the data
2. Why the survey is designed as it is
3. The complications inherent in this and similar studies
4. The assumptions necessary for a workable analysis of the data, and
5. The formulas necessary for the computation of the point estimates and their variances.

The following section is primarily to explain the logic behind the sampling scheme and may be skipped by those interested only in the actual study. Necessary information will be repeated in subsequent sections.

PRELIMINARY CONSIDERATIONS

Due to cost factors, the survey must be designed in such a way that one man working regular eight-hour days and an average of a forty hour week will be able to collect all the necessary data. Thus, we have about 50 work days during the ten week period which are to be used in such a way as to maximize the precision of our estimates.

Sufficient data are to be collected to yield a coefficient of variation of 5% or less on all the estimates (up to the limits imposed by the restriction on labor mentioned above).

The requirement of an eight hour work day precludes the possibility of randomly sampling units of time at which data might be collected and requires that we take "cluster samples" of time where each cluster is approximately

eight hours in length. With this end in mind, each day of the season is considered as a primary cluster and is divided into two secondary clusters of $8\frac{1}{2}$ hours each (the period from 5 a.m. to 1:30 p.m. and the period from 1:30 p.m. to 10 p.m.). The approach is to draw a random sample of fifty days (primary clusters) and to randomly select one of the work periods (a secondary cluster) within each day, thus determining the fifty actual work periods during which data are to be collected. (An $8\frac{1}{2}$ hour designated work period is felt to be acceptable since the employee will not, generally, work continuously.)

There are two sources of variation for any estimate derived from cluster sampling: variance from within the clusters and variance between clusters. Little can be done to control variance within clusters. It is, however, possible to reduce the variance between clusters through stratification of the season into segments which are as alike as possible and, as will be seen later, it is essential that we use every method available to reduce this component of variation. Therefore, it is necessary to stratify the days of the season into the three relatively homogeneous strata listed below.

Stratum I: Opening Day

Stratum II: All other weekend days plus New Years and Lincoln's Birthday

Stratum III: All week days except New Years and Lincoln's Birthday

Since stratum I consists of only one day, it is necessary to take the somewhat dubious approach of "collect all the data possible and hope for the best." Four men will work continuously throughout the day and all data obtained will be assumed to represent a random sample from this stratum.

Physical limitations prevent adopting a more sophisticated approach to this stratum due to its size and, yet, it is felt that it differs significantly enough from the other strata to preclude merging it with either.

A portion of the data collected on Georgetown Lake during the 1968-69 ice fishing season was used to estimate the approximate variances which we might expect in this survey and, using these estimates, the information required in Strata II and III was determined to be as follows:

Stratum II: Interviews with about 700 fishermen and counts of the number of cars on the lake at 168 different times.

Stratum III: Interviews with about 500 fishermen and counts of the number of cars on the lake at 232 different times.

We assumed that not more than an average of one car count per hour should be scheduled which implies that the interviewer should work 21 days in Stratum II and 29 days in Stratum III. Thus, all of the days in Stratum II are to be worked and 29 of the 48 days in Stratum III are to be randomly selected as work days.

The complete process for selecting the sample is outlined in the following section.

SAMPLING PROCEDURE

I. STRATIFY THE SEASON

A. Stratum I: Opening day

1. The small size of this stratum requires that intensive effort be expended to collect all possible data on both car counts and fishermen. Four men will work throughout the day and all data collected will be assumed to represent a random sample.

B. Stratum II: Weekend days plus New Years and Lincoln's Birthday




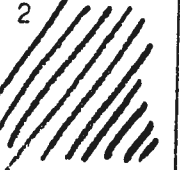




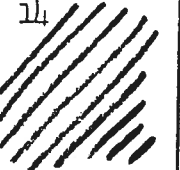
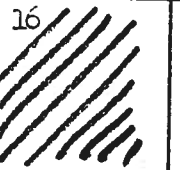

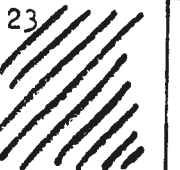
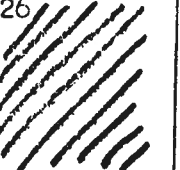

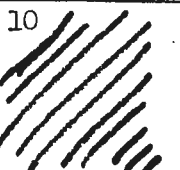
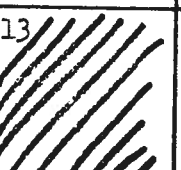
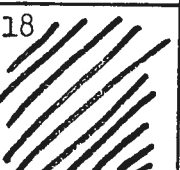
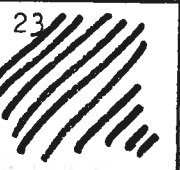
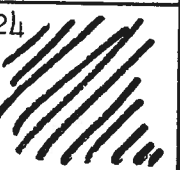
1. All 21 days are to be worked.
2. In each day, randomly determine if the interviewer will work the a.m. shift (5 a.m. to 1:30 p.m.) or the p.m. shift (1:30 p.m. to 10 p.m.).
3. Randomly select 168 times at which car counts are to be made during these work periods.
4. Fishermen leaving the lake are to be contacted as completely as possible while adhering to the car count schedule. They are to be asked the total number of fish caught, the total number of hours fished, and the number of fishermen who arrived in the same automobile as they did. This information is to be gathered on a per man basis. These contacts are assumed to represent a random sample of the men fishing that day.

C. Stratum III: All week days except New Years and Lincoln's Birthday

1. Christmas day is included in this stratum since it is felt that with respect to fishing activity it is more like a week day than a week end day. The interviewer will not be asked to work Christmas day although it is treated in our analysis as though it could have fallen into our sample.
2. Randomly select 29 of the remaining 47 days in this stratum.
3. Randomly determine if the interviewer is to work the a.m. shift or the p.m. shift in each of the days selected.
4. Randomly select 232 times at which car counts are to be made during these work periods.
5. Fishermen leaving the lake are to be contacted as completely as possible while adhering to the car count schedule. The data collected is the same as that in stratum II. The fishermen contacts are assumed to represent a random sample of the men fishing.

The calendar on the following page shows the work periods determined through the first two stages of sampling (selection of primary clusters and selection of secondary clusters). Days off are blacked out and the work days are designated as a.m. or p.m. depending upon which shift the interviewer is to work. The list of the scheduled car counts (third stage of sampling) is too extensive to be included on the calendar. It is, however, on file for future reference if that is necessary.

GEORGETOWN LAKE WINTER CREEL CENSUS SCHEDULE

| Sunday | Monday | Tuesday | wednesday | Thursday | Friday | Saturday |
|-----------------------|---|---|---|--|--|----------|
| 21 AM and PM | 22 AM | 23 AM | 24  | 25  | 26 PM | 27 AM |
| 28 PM | 29  | 30 AM | 31 AM | 1 PM | 2  | 3 AM |
| 4 AM | 5 AM | 6  | 7  | 8 AM | 9  | 10 PM |
| 11 AM | 12  | 13 AM | 14  | 15 AM | 16  | 17 AM |
| 18 PM | 19 PM | 20 AM | 21 AM | 22  | 23  | 24 AM |
| 25 PM | 26  | 27 PM | 28 PM | 29 PM | 30 PM | 31 PM |
| 1 AM | 2 PM | 3 PM | 4 PM | 5  | 6 AM | 7 PM |
| 8 AM | 9 PM | 10  | 11 AM | 12 PM | 13  | 14 AM |
| 15 AM | 16 PM | 17 PM | 18  | 19 PM | 20 PM | 21 AM |
| 22 PM | 23  | 24  | 25 AM | 26 AM | 27 AM | 28 PM |

COMPLICATIONS

Since the sampling is done in three stages, the data should be analyzed using techniques related to multi-stage cluster samples. It is, however, necessary to deviate from the standard approach since it requires that the size of the population (number of fishermen) be known a priori for each of the ultimate clusters (work days). This is one of the variables to be estimated from the survey so it is not possible to adhere to standard practices. We are faced with three alternative approaches to this problem:

- (1) Estimate totals for each work day and attempt to derive the necessary variances on the basis of theoretical statistics.
- (2) Estimate the totals for each work day and treat them as constants for use in the usual computational formulas, or
- (3) Ignore the fact that we have a three stage sampling design and assume that the final sample data is identical to that which would have been obtained if a simple random sample were possible. (This is equivalent to assuming that there is no variance between ultimate clusters in any one stratum so the prudence of the stratification employed should be apparent.)

The colossal amount of work involved in approach (1) makes it unacceptable. Three stage sampling designs have nearly intractable computations associated with them under the best of conditions and we would have to introduce a different random variable for each day thus expanding the equations to impossible lengths.

Approach (2), although being attractive since it allows us to use standard three stage formulas, could only be used with a great deal of

caution since it is not intuitively clear what effect it would have on the estimates. It is certainly no more acceptable than approach (3) and its use might quite possibly introduce unmeasurable biases into the point estimates.

With these considerations in mind, it was decided to use approach (3). This is not too unacceptable in view of the fact that our sample can be viewed under any conditions as an "equal probability sample" (i.e., the chances of a fisherman being interviewed are about proportional to the number of fishing trips he makes to the lake) and therefore there is no bias introduced into our point estimates through this approach. The estimated variances will be biased somewhat since it is probably not true that the variance between clusters is zero. However, the strata have been set up in such a way as to minimize this component of the total variance within any one stratum.

These comments have been included to indicate that full consideration has been given to the "preferable" approach even though it is intractable due to complications inherent in the study. The equations in the formulas section of this paper require only a minute fraction of the labor which would be involved in either approach (1) or (2) and it will be noted that they are themselves somewhat complex.

ASSUMPTIONS

The additional assumptions made in the analysis of the data from the survey are stated here since the formulas involved are such that these assumptions may not be readily apparent to the casual reader. If for no other reason, they are stated to forestall possible criticism arising from their omission.

The final estimate of each of the variables under consideration will generally be the product or quotient of several random variables. In calculating the variance of such an estimate, it is necessary to have some knowledge of the covariance (degree of relationship) between the variables involved. If there is no relationship then the formula for the variance is somewhat simplified whereas if there is some relationship, the covariance of the variables must be estimated from the sample data and incorporated into the formula. It is fortunate that we are able to reasonably assume independence (no relationship) in the case of all but one pair of random variables. Specifically, we will assume that:

- (1) Size of fishing party is independent of the car count
- (2) Size of fishing party is independent of hours fished per man
- (3) Size of fishing party is independent of fish caught per man
- (4) The car count is independent of hours fished per man, and
- (5) The car count is independent of fish caught per man.

≡ in our case, fishermen per boat is independent of boat count.

The assumption in each case is equivalent to assuming that the covariance between the two variables is zero. For example, assumption (5)

could be concisely stated as $COV(C,H) = 0$. Simply stated, we are assuming that knowledge of one variable yields no information about the other.

It is felt that no notable error is introduced through making these assumptions. It is, of course, unreasonable to assume that fish caught is independent of hours fished and the covariance of these two variables must be accounted for in the estimate of the variance of their quotient (fish per hour).

NOTATION

For the sake of simplicity, the estimation formulas will be developed as though there were only one stratum and the notation is developed accordingly. At the end of the formulas section, the methods for deriving the estimates for the entire season will be given.

M : Total number of hours in the season

m : Number of car counts made during the season

c_i : Number of cars counted on the i 'th count; $i = 1, 2, \dots, m$

TFT : Total number of fisherman trips made during the season

\bar{f}_m : Average number of men fishing at a random point in time

n : Number of fishermen contacted during the season

f_i : Number of fish taken by the i 'th fisherman interviewed; $i=1, \dots, n$

h_i : Number of hours fished by the i 'th fisherman interviewed; $i=1, \dots, n$

m_i^C : Number of men who arrived in the same car as the fisherman being interviewed for the i 'th contact; $i=1, \dots, n$.

TMH : Total man hours expended on the lake during the season

TFH : Total fish harvest during the season

f/h : Average fish per hour during the season

VAR(.) : Denotes the variance of the variable within the parenthesis

COV(.,.) : Denotes the covariance of the 2 variables within the parenthesis

$\bar{\cdot}$: Denotes the average of the variable considered (a variable with a bar over it will always denote the average over all sampled units)

$\sum_{i=j}^k x_i$: Denotes the sum of the subscripted variable over all subscripts ranging from j to k .

FORMULAS

$$(1) \bar{c} = \frac{\sum_{i=1}^m c_i}{m}$$

$$(2) \text{VAR}(\bar{c}) = \frac{\sum_{i=1}^m (c_i - \bar{c})^2}{m(m-1)}$$

In order to estimate [average men per car] it is necessary to convert the data from a per man basis to a per car basis. This is done through the use of weighting factors which offset the varying probability of a car being represented through the sample of fishermen. (The probability of contacting at least one man in a two man party is twice as great as for a one man party, so this must be adjusted for in order to yield an unbiased estimate of men per car.) The two formulas given below satisfy our needs.

$$(3) \bar{m}^c = \frac{n}{\sum_{i=1}^n \frac{1}{m_i^c}}$$

has got to be some form of unbiased estimate of the mean for sampling with probability proportional to size

$$(4) \text{VAR}(\bar{m}^c) = \frac{\sum_{i=1}^n \left\{ \frac{1}{m_i^c} (m_i^c - \bar{m}^c)^2 \right\}}{\sum_{i=1}^n \left\{ \frac{1}{m_i^c} \right\} \left[\sum_{i=1}^n \left\{ \frac{1}{m_i^c} \right\} - 1 \right]}$$

$$(5) \bar{f}_m = \bar{m}^c \cdot \bar{c}$$

[Since \bar{m}^c is independent of \bar{c} we can estimate:]

$$(6) \text{VAR}(\bar{f}_m) = (\bar{m}^c)^2 \text{VAR}(\bar{c}) + (\bar{c})^2 \text{VAR}(\bar{m}^c)$$

$$(7) \text{TMH} = M * \bar{f}_m$$

$$(8) \text{VAR}(\text{TMH}) = M^2 \cdot \text{VAR}(\overline{f_m})$$

$$(9) \overline{h} = \frac{\sum_{i=1}^n h_i}{n}$$

$$(10) \text{VAR}(\overline{h}) = \frac{\sum_{i=1}^n (h_i - \overline{h})^2}{n(n-1)}$$

$$(11) \text{TFT} = \frac{\text{TMH}}{\overline{h}}$$

[Since \overline{M} , \overline{c} , and \overline{h} are independent of one another, we can estimate:]

$$(12) \text{VAR}(\text{TFT}) = \frac{(\text{TMH})^2}{(\overline{h})^2} \left\{ \frac{\text{VAR}(\text{TMH})}{(\text{TMH})^2} + \frac{\text{VAR}(\overline{h})}{\overline{h}^2} \right\}$$

$$(13) \overline{f} = \frac{\sum_{i=1}^n f_i}{n}$$

$$(14) \text{VAR}(\overline{f}) = \frac{\sum_{i=1}^n (f_i - \overline{f})^2}{n(n-1)}$$

$$(15) (f/h) = \frac{\overline{f}}{\overline{h}}$$

Estimate (15) is the quotient of two random variables which are not independent so it is necessary to account for their covariance in the estimate of the variance.

$$\begin{aligned}
(16) \text{VAR}(f/h) &= \frac{\bar{f}^2}{\bar{h}^2} \left\{ \frac{\text{VAR}(\bar{f})}{\bar{f}^2} + \frac{\text{VAR}(\bar{h})}{\bar{h}^2} - 2 \frac{\text{COV}(\bar{f}, \bar{h})}{\bar{f} \cdot \bar{h}} \right\} \\
&= \frac{\bar{f}^2}{\bar{h}^2} \left\{ \frac{\text{VAR}(\bar{f})}{\bar{f}^2} + \frac{\text{VAR}(\bar{h})}{\bar{h}^2} - 2 \frac{\sum_{i=1}^n (f_i - \bar{f})(h_i - \bar{h})}{n(n-1)} \right\} \\
&= \frac{1}{\bar{h}^2} \left\{ \frac{\sum_{i=1}^n (f_i - \frac{\bar{f}}{\bar{h}} h_i)^2}{n(n-1)} \right\}
\end{aligned}$$

$$(17) \text{TFH} = (f/h) \cdot \text{TMH}$$

There are four random variables implicit in formula (17) - \bar{f} , \bar{h} , \bar{M}^c , and \bar{c} . However, they are all independent of one another so we may estimate:

$$(18) \text{VAR}(\text{TFH}) = (f/h)^2 \text{VAR}(\text{TMH}) + (\text{TMH})^2 \text{VAR}(f/h)$$

These eighteen equations will yield the estimates and each of their variances for any one stratum. To indicate the procedure for deriving estimates pertinent to the entire season, assume that everything in equations (1) through (18) has an additional subscript, say k , which denotes the stratum under consideration. $k = 1, 2, \text{ or } 3$.

The totals can be derived as the sum of the stratum totals and the variance of such a total is, due to the nature of stratified sampling,

simply the sum of the variances from the three strata. Specifically, we have

$$(19) \quad TMH = \sum_{k=1}^3 (TMH)_k$$

$$(20) \quad VAR(TMH) = \sum_{k=1}^3 VAR([TMH]_k)$$

$$(21) \quad TFT = \sum_{k=1}^3 (TFT)_k$$

$$(22) \quad VAR(TFT) = \sum_{k=1}^3 VAR([TFT]_k)$$

$$(23) \quad TFH = \sum_{k=1}^3 (TFH)_k$$

$$(24) \quad VAR(TFH) = \sum_{k=1}^3 VAR([TFH]_k)$$

An average for the entire season is derived as the weighted average of the stratum averages where the weight assigned to any stratum is proportionate to a measure of its respective size. In deriving the

variance of these weighted averages, the weights are treated as constants. Specifically, we have:

$$(25) \quad \bar{f} = \sum_{k=1}^3 \left\{ \frac{(TFT)_k}{TFT} \cdot \bar{f}_k \right\}$$

$$(26) \quad VAR(\bar{f}) = \sum_{k=1}^3 \left\{ \frac{(TFT_k)^2}{(TFT)^2} VAR(\bar{f}_k) \right\}$$

$$(27) \quad \bar{h} = \sum_{k=1}^3 \left\{ \frac{(TFT)_k}{TFT} \cdot \bar{h}_k \right\}$$

$$(28) \quad VAR(\bar{h}) = \sum_{k=1}^3 \left\{ \frac{(TFT_k)^2}{(TFT)^2} VAR(\bar{h}_k) \right\}$$

The average fish per hour is the ratio of two random variables and cannot be considered as an average over sampling units. It is known as a "ratio estimate" and has unavoidable biases built into it. There are several reasonable ways of estimating a ratio pertinent to the entire season of which one is to simply take a weighted average as we did for fish per man and hours per man. It is possible, however, for the biases within the strata (which are assumed to be negligible) to accumulate into a non-negligible bias in the final estimate using this approach. This

tendency can be avoided by using the estimated averages pertinent to the entire season to derive our ratio estimate for the entire season and, although the formula for the variance requires more work, this approach is taken. (This approach also forces the three final averages to be consistent among themselves which is not a characteristic of the weighted averages approach.) Thus, our estimate is derived for the season exactly as it was derived within each stratum.

$$(29) \quad (\bar{f}/\bar{h}) = \frac{\bar{f}}{\bar{h}}$$

This is, again, the quotient of two random variables which are not independent so it is necessary to estimate their covariance from the stratified sample. This can be estimated as:

$$(30) \quad \text{COV}(\bar{f}, \bar{h}) = \sum_{k=1}^3 \left\{ \left(\frac{TFT_k}{TFT} \right)^2 \text{COV}(\bar{f}_k, \bar{h}_k) \right\}$$

Which can be estimated since within the kth stratum we have (as given in equation (16))

$$(31) \quad \text{COV}(\bar{f}_k, \bar{h}_k) = \frac{\sum_{i=1}^{n_k} (f_{ki} - \bar{f}_k)(h_{ki} - \bar{h}_k)}{n_k(n_k - 1)}$$

Thus, using equation (30) and (31), we are finally able to estimate

$$(32) \quad \text{VAR}(\bar{f}/\bar{h}) = \frac{(\bar{f})^2}{(\bar{h})^2} \cdot \left\{ \frac{\text{VAR}(\bar{f})}{(\bar{f})^2} + \frac{\text{VAR}(\bar{h})}{(\bar{h})^2} - 2 \frac{\text{COV}(\bar{f}, \bar{h})}{\bar{f} \cdot \bar{h}} \right\}$$

The standard error of estimate can be derived in all cases as the square root of the variance. Approximate 95% confidence limits can be computed as the point estimate plus or minus two standard errors.