SAMPLING DESIGN FOR MARINE RECREATIONAL FISHING SURVEY: AN EXAMPLE OF A COMPLEMENTED SURVEY

Dhirendra N. Ghosh, Consultant, Washington, D.C.

The U.S. Federal government has long maintained detailed statistics concerning commercial saltwater finfish catch. Information on recreational catch was difficult to obtain. The major obstacles were the estimation of participants and access points and chronic inability of recreational fishermen to identify the species of the fish they catch as well as their short memory regarding the number and weight of fish caught in a fishing trip undertaken, even a few days back.

This complemented survey methodology grew out of a contract awarded to Human Sciences Research, Inc., by the National Marine Fisheries Service in 1976. It solves the problems mentioned above. Moreover, this methodology can also be successfully applied to various other situations we will describe below.

The problem is to estimate the total number and weight of fish caught from marine waters by each species by all recreational fishermen during a period of time, and the number of participants in recreational fishing.

The usual survey methodology of selecting a random sample of households and collecting data on marine recreational fishing activities of all members of the households and collecting data on marine recreational fishing activities for all members of the household was tried several times prior to the award of the above mentioned contract but failed for the following reasons.

1. Fishermen can scarcely identify species that are rather uncommon. Their performance even with common species is not very commendable.
2. Their specieswise estimate of weight and number of fish caught even the day before are highly erroneous.
3. It is not cost-effective to sample households even 50 miles from the coast since such households rarely report saltwater finfishing trips.

The complemented survey is basically two independent surveys to estimate variables that are later combined to estimate the desired variable. In this case, two surveys are undertaken, a survey of households in the coastal areas to estimate the total number of saltwater finfish trips undertaken by members of the sample during a period of two months. In this series of surveys being conducted by NMFS since the development of this methodology, this household survey has been a telephone survey using the Random Digit Dialing Technique. The coastal zone is stratified into a number of regions and simple random sample of households are selected from each county that is situated within a certain number of miles from the coast. The total number of trips is broken down by four modes of fishing:

1. Beach or bank;
2. manmade structure;
3. private and rental boats; and
4. party and charter boats.

The complimentary survey during the same time period is a sample survey of fisheries on the beach after their fishing trip is completed. Their catch is counted and weighed and species identified by trained interviewers. The objective of this complimentary survey is to estimate the mean catch per fishing trip for each of the four modes of fishing mentioned above. If the average catch per trip is multiplied by the independent estimate of the total number of fishing trips undertaken during a period of time the estimate of the total number of trips undertaken by fishermen living in the noncoastal zone. If the average catch of coastal fishermen differ from noncoastal fishermen two separate estimates of mean catch are used to multiply the corresponding estimates of the total number of trips.

The sampling frame for the intercept survey is the list of all fishing sites. These fishing sites are first stratified by mode of fishing and by state. On any particular day, a number of fishermen fish at a site. Accordingly, a site constitutes a cluster of fishermen. This constitutes an Epsem (equal probability selection method) sample. However, this is not a simple random sample of fishermen even though the probability of selection is equal for every fisherman. In sampling theory, when a mean is to be estimated, there is no need to know the individual probabilities of selection. The only requirement is that the probabilities are equal.

This method presupposes that individual fishermen are interviewed at the completion of their trips because otherwise the probabilities of selection will not remain equal. However, this type of interviewing is very expensive when only a few fishermen are fishing at a site and the interviewer has to remain at the site until all the fishermen leave. Thus, the per-unit cost of collecting data is much higher in low pressure sites than in high pressure sites.

An alternative is to resort to collecting data from fishermen still fishing. However, there are two problems associated with this method of data collection. First, in order to estimate total catch, we need the total catch per sampled trip. Fishing *trip* is the unit of analysis. If we find out the time the fisherman has already spent fishing at the time of the interview and how much longer he intends to fish after the interview, the total catch at the time of interview can then be properly inflated to obtain an estimate of the total catch for any fisherman who was interviewed using this method. If his estimate of the time he has already been fishing and the time he intends to stay after the interview are not biased in any direction, upward or downward, the inflated estimate will have a larger sampling error but no systematic bias.

There is a second problem, however. A fisherman who fishes a longer time will have a larger probability of being selected in the sample if data are collected from fishermen while they are still fishing. This, of course, must be taken into account when the estimates are built.

This type of sampling is known as "size biased sampling."[1] This type of sampling falls into the broad class of sampling known as "probability proportional to size" (PPS). The theory is as follows. Assume that $x_1 x_2 \ldots x_n$ are the values obtained from a sample such that the probability of selection of each x_i is proportional to its size. If we calculate the arithmetic mean, that mean will be biased toward large values since they had a higher probability of selection. It can be shown that the unbiased estimate of the mean is:

$$\left[\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_n}\right)\right]^{-1}$$

This is in essence the harmonic mean of the x's.

As an example, if a fisherman had fished for six hours when he was intercepted, and he was going to fish for another four hours, his probability of selection is proportional to ten hours. If M is the average duration of a fisherman's trip, his probability of selection is $\frac{10}{NM}$, (where N is the total number of fishermen and NM stands for the total fishing time of all fishermen).

If "a" stands for the catch of the fisherman when he is contacted, then his inflated catch is $\frac{a}{6} \times 10$.

From the theory of PPS sampling, we know $\frac{xi}{pi}$ is an unbiased estimate of the total, if xi is the value of the variable and pi is its probability of selection, therefore:

$$\frac{10}{6} \div \frac{10}{NM} = \frac{a}{6} NM$$

is an unbiased estimate of the total catch of all fishermen. Consequently, $\frac{a}{6}$ M is an unbiased estimate of the average catch.

M (fishing trip length) can be estimated by two different ways. (1) Calculating the narmonic mean of the total fishing time of fishermen in the sample, or (2) calculating the mean time from the trips in the same mode and state for all the intercepts which were "complete trip" intercepts. A combination of the above two can be used if the sample size is small for both types of interviews.

For example, assume we interview two fishermen while they fish, and that their total fishing time (past plus anticipated) is 10 and 6 hours, respectively. The arithmetic mean of 10 and 6 is 8, but the harmonic mean is 7.5. The estimate for the mean fishing time from the sample should, therefore, be 7.5, and not 8.

Estimation of Participants

This is perhaps the most complex estimation component of the whole survey. When a fisherman is intercepted, he cannot know the total number of trips he will undertake during the calendar year. As a substitute, he is asked to report the total number of trips he undertook during the past year from the day of the intercept. If we designate the number of trips he reports as K, we assume that the total number of trips he will make in the calendar year, the reference period of the survey, is also K. We then associate the fraction $\frac{1}{K}$ to the trip intercepted. Therefore, to every trip he has taken, or will take, one such fraction can be associated. Obviously, the sum of all these fractions over all the trips undertaken in a state in a year is the total number of participants we want to estimate.

The sample provides an estimate of the mean of these fractions. By multiplying this by the estimate of total trips, we can estimate the sum of all these fractions. From the point of view of sampling theory, a fisherman making K trips in a year has a probability of selection K times the probability of selection of a fisherman making only one trip. This is another example of probability proportional to size (PPS) sampling. If we count each intercepted fisherman as one participant, then he will be counted K times over the year. By counting him as $\frac{1}{K}$ we rectify for the higher probability of his selection.

Estimation of Sampling Variance

No scientifically conducted survey is complete without the calculation of the measure of precision of the estimates involved. The complemented surveys methodology employed in this study is complex, and the calculation of the measures of precision is equally complex. The following discussion presents the basic theory and the procedures involved.

Intercept Survey

There are two choices for the estimation of the sampling variance for the intercept survey. One is the ratio-to-size estimate for cluster samples and the other is to treat the sample as a simple random sample. Even though the estimates are the same, their sampling variances are different.

In the case of cluster sampling, the size of the cluster is treated as the auxiliary variable and the formula follows for the standard formula for sampling variance for ratio estimates.

$$\hat{V}(\hat{\bar{Y}}_R) = \frac{\sum_{i=1}^{n} M_i^2 (\bar{yi} - \hat{\bar{Y}}_R)^{2^2}}{n\bar{M}^2 (n-1)}$$

Where: yi = mean for the ith cluster in the sample

M_i = size of the ith cluster in the sample

$\bar{M} = \frac{\sum_{i=1}^{n} M_i}{n}$

n = number of clusters in the sample

$\hat{\bar{Y}}_R$ = the estimate for the mean

However, if the size of the cluster is independent of the mean of the cluster, the formula for simple random sampling provides a close approximation, and is much simpler to calculate. From our past experience, we have found this to be true. In other words, it was found that the clustering effect is negligible.

Complemented Surveys

The estimation of sampling variance for the combined estimates for the two surveys, however, requires special attention. The intercept survey provides estimates (x) for the average catch per trip for a mode within a state. The sampling variance v(x) is estimated from the sample as described above. Let w be the estimate of the total number of trips from the telephone zone in the state for the corresponding mode, obtained from the household survey. Let p be the proportion of trips from the telephone zone to the total trips intercepted. Then $y = \frac{w}{p}$ is the estimate for the total number of trips within a state for a mode.

$y = \frac{w}{p}$ is a ratio of two independent random variables and its sampling variance is estimated as follows:

$$\frac{V(y)}{[E(y)]^2} = \frac{V\frac{w}{p}}{[E(\frac{w}{p})]^2} = \frac{V(w)}{(Ew)^2} + \frac{V(p)}{(Ep)^2}$$

where E is the expected value:

$$v\left(\frac{w}{p}\right) = \left(\frac{w}{p}\right)^2 \left[\frac{v(w)}{w^2} + \frac{v(p)}{p^2}\right]$$

is a consistent estimate of V(y).

This follows because of a theorem in sampling theory "which states that a rational function of consistent estimates is a consistent estimate of the same rational function of the quantities being estimated, if the denominator does not vanish when the quantities estimated are substituted in the rational function."[3]

Let z = xy be the estimate for the total fish caught (for individual species and also all species combined). It is possible to calculate the sampling variance of z in terms of the expected values and sampling variance of x and y.[4] This can also be estimated from the two samples.[5] The estimation of the sampling variance for participants is a direct application of these formulas.

The central limit theorem assures us that both x and y are approximately normally distributed for large sample sizes. However, z is not[6] normally distributed, its statistical distribution is $f(u) = \frac{1}{\pi} \int_0^{\infty} (1 + t^2)^{-1/2} \cos(ut) = \frac{1}{\pi} K_0(u)$ where K_0 is the modified Bessel function of the third kind.

The algebra is extremely complex for calculating the usual 95 percent confidence intervals for the final estimate. Hence, we performed several "Monte Carlo" experiments with several thousands of standard, normal deviates for the construction of empirical distributions with different hypothetical values of means and standard deviations. It was determined that the normal approximation works very satisfactorily for constructing the interval for 95 percent confidence.

In the actual survey, the whole population is divided into a large number of strata. Even though the distribution of z in each stratum is unknown, when a variable is summed over a large number of strata, the central limit theorem of mathematical statistics assures us of the validity of the normal approximation. Thus, both theoretical reasoning and experimental evidence indicates that normal approximation would work satisfactorily.

Optimum Allocation

Given an established acceptable level of relative variance for the final estimate, it is possible to compute the sample sizes for the two surveys in such a way that total costs are minimized. If we establish the requirement that the relative variance of the estimate must be equal to or less than some stated level, the optimal sample sizes can be calculated as follows:

Let C_0 = fixed cost
C_1 = cost of contacting each fisherman in the intercept survey
C_2 = cost for contacting each household in the telephone survey
K^2 = the relative variance of the final estimate
K_1, K_2 = coefficients of variation of the two populations
n_1 = optimal sample size for the intercept survey
n_2 = optimal sample size for the telephone survey

The total cost of conducting both surveys is then equal to:

$$C_0 + C_1 n_1 + C_2 n_2$$

The Lagrangian Multiplier technique can then be used to determine the optimal values for the two sample sizes n_1 and n_2 as follows:

$$n_1 = \frac{1}{K^2}\left[\frac{\sqrt{C_2}}{\sqrt{C_1}} K_1 \cdot K_2 + K_1^2\right], \text{ and}$$

$$n_2 = \frac{1}{K^2}\left[\frac{\sqrt{C_1}}{\sqrt{C_2}} K_1 \cdot K_2 + K_2^2\right]$$

it is necessary to insure that the sample size for each survey is large enough to provide estimates for the parameters with sufficient precision.

---

[1]G. P. Patie and J. K. Ord, "On Size-Biased Sampling and Related Form-Invariant Weighted Distributions," *Sankhya, 38*, Series B. Pt. 1 (1976), pp. 48-61. R. L. Scheaffer, "Sized-Biased Sampling," in *Technometrics, 14* (1972), pp. 635-644.

[2]William Cochran, *Sampling Techniques,* John Wiley & Sons, 3rd Edition, 1977, page 250.

[3]Mr. Hansen, W. Hurwitz, and M. Madow, *Sample Survey Methods and Theory, Volume II,* John Wiley and Sons, 1953, page 120.

[4]The formula is: $V(z) = V(x) \cdot V(y) + [E(x)]^2 \cdot (V(y) + [E(y)]^2 \cdot V(x)$

[5]The formula is: $v(z) = x^2 v(y) + y^2 v(x) - v(x) \cdot v(y)$ where v(x) is an unbiased estimate of V(x) and v(y) is an unbiased estimate of V(y). Source: Leo Goodman, "On the Exact Variance of Products," *Journal of the American Statistical Association, 55,* (1960) pp. 708-713.

[6]Sources: M. A. Kendall and A. Stuart, *The Advanced Theory of Statistics,* Volume I (New York: Hafner Publishing Company, 1969). G. N. Watson, *Theory of Bessel Functions* (Cambridge: Cambridge University Press, 1962).