2005

Gulf Coast Alaska Fisheries Economic Activity Survey

SAMPLING PROCEDURES FOR HARVESTING SECTORS[1]

The overall project objective is to estimate the employment and labor income information for each of three disaggregated harvesting sectors using data to be collected via a mail survey. Using ex-vessel revenue information, an unequal probability sampling (UPS) procedure will be employed to determine the sampling plan for each of the three harvesting sectors. The procedure is described below.

In the literature, there exist many methods for conducting UPS without replacement (see, for example, Brewer and Hanif 1983; Sarndal 1992). One critical weakness with most of these methods is that the variance estimation is very difficult because the structure of the 2nd order inclusion probabilities (π_{ij})[2] is complicated. One method that overcomes this problem is Poisson sampling. However, one problem with Poisson sampling is that the sample size is a random variable, which increases the variability of the estimates produced. An alternative method that is similar to Poisson sampling but overcomes the weakness of the Poisson sampling is Pareto sampling (Rosen 1997)[3] which yields a fixed sample size.

In this project, there are two tasks that we need to do for estimating the population parameters using UPS without replacement. First, the optimal sample size needs to be determined. Second, once the optimal sample size is determined, the population parameters and confidence intervals need to be estimated. For the first task, we will use the variance of Horvitz-Thompson (HT) estimator from Poisson sampling in Part I below.[4] For the second task, we will use the Pareto sampling method described in Part II below (Slanta 2006). In determining the optimal sample size in Part I, we will use information on an auxiliary variable (ex-vessel revenue). To estimate the population parameters in Part II, we use actual response sample information on the variables of interest (employment and labor income).

Part I: Estimating Sample Size

Step 1: Estimation of Optimal Sample Size (n*)

(A) Obtaining Initial Probabilities

To obtain the initial values of the inclusion probabilities (π_i) for unit i in the population, we multiply the auxiliary value of unit i (X_i, i.e., the ex-vessel value of vessel i in the population) by a proportionality constant (t)[5]:

$$\pi_i = t X_i \qquad (1)$$

where π_i : probability of vessel i being included in the survey sample
X_i : value of the auxiliary variable (ex-vessel value of vessel i in the population)

Here, t is given by

$$t = \frac{\sum_i^N X_i}{V + \sum_i^N X_i^2} \qquad (2)$$

where N : population size
V : desired variance (of HT estimator of the population total); Poisson variance. Here, V is given as:

$$V = \left(\frac{\varepsilon X}{z_{1-(\alpha/2)}} \right)^2$$

where ε is the error allowed by the investigator [e.g., if ε is 0.1, then 10% error of true population total ($X = \sum_{i=1}^N X_i$) is allowed]; and z is percentile of the standard normal distribution. Therefore, choosing a desired variance V is equivalent to setting the values of ε and z. The value of V calculated using $V = \sum_{i=1}^N \frac{(1-\pi_i)X_i^2}{\pi_i}$ (Poisson variance; Brewer and Hanif 1983, page 82) with π_i's being the final values of N inclusion probabilities obtained from Step 1, will be equal to the desired variance given at the beginning of Step 1.

Some of the resulting π_i's could be larger than one. The number of certainty units (i.e., the number of units for which $\pi_i > 1$) is denoted C_1. If $\pi_i > 1$, then we force this inclusion probability to equal one ($\pi_i = 1$).

(B) Iterations and Determination of Optimal Sample Size

We recalculate t using the noncertainty units (i.e., the units for which $\pi_i < 1$) obtained in (A) above, i.e.,

$$t = \frac{\sum_{i}^{M_1} X_i}{V + \sum_{i}^{M_1} X_i^2} \tag{2'}$$

where M_1 : number of noncertainty units from (A), where $M_1 = N - C_1$.

Using equation (1) above, we calculate the inclusion probabilities for the noncertainty units by multiplying the t value [from equation (2')] by the ex-vessel values of the noncertainty units. If the resulting π_i's are larger than one, we force them to equal one. The resulting numbers of certainty and noncertainty units are denoted C_2 (= C_1 + additional number of certainty units) and M_2 (= M_1 – additional number of certainty units), respectively, where $C_2 + M_2 = N$. Next, for M_2 units of noncertainty, we calculate the t and π_i's again. This is an iterative process. We continue this process until the noncertainty population stabilizes (i.e., until there is no additional certainty unit).

If the noncertainty population stabilizes after *k*th iteration, there will be C_k units of certainty units and M_k units of noncertainty units and $C_k + M_k = N$. Summing over the probabilities for all these certainty and noncertainty units, we obtain the optimal sample size (n*) as:

$$n^* = \sum_{i}^{N} \pi_i \tag{3}$$

At this stage the optimal sample size may not be an integer number. In this stage, we also compute the optimal sample size under simple random sampling (SRS)[6], n_{srs}, and compare it with n*.

<u>Step 2: Determining Number of Mailout Surveys</u>

(A) Adjustment of Probabilities

Once the optimal sample size (n*) is determined in Step 1, we divide the sample size (n*) by the expected response rate (obtained from previous studies) to determine the number of surveys that need to be mailed out to achieve n*. The number thus derived is denoted n_a (this number may not still be an integer value). We next adjust the inclusion probabilities for the M_k noncertainty units obtained in Step 1 above as:

$$\pi_i = (n_a - C_k)\left[\frac{\pi_i}{\sum_{i}^{M_k} \pi_i}\right] \tag{4}$$

If the resulting probabilities are larger than one ($\pi_i > 1$), we make them certainties ($\pi_i = 1$). The resulting numbers of certainty and noncertainty units are denoted C_{k+1} and M_{k+1}, respectively. Next, we adjust the probabilities of the new set of noncertainty units (M_{k+1}) in a similar way using equation (4') below:

$$\pi_i = (n_a - C_{k+1})\left[\frac{\pi_i}{\sum_i^{M_{k+1}} \pi_i}\right] \qquad (4')$$

We continue this process until the noncertainty population stabilizes. The resulting numbers of certainty and noncertainty units are C_q and M_q, respectively.

(B) Apply Minimum Probability Rule

At this point, we impose a minimum probability rule. UPS can have excessively large weights ($= 1/\pi_i$) and if they report a large value, then the population estimate and its variance would be very large. In order to avoid this problem, we can impose a minimum value of the inclusion probabilities. If m is the minimum imposed probability, then we do the following:

If $\pi_i < m$, then set $\pi_i = m$ for each i, where $i = 1, \ldots, N$.

The value for m here is determined arbitrarily. The only cost involved in using this rule is a small increase in sample size.[7]

(C) Finding an Integer Value for Sample Size

Next, we add up all the resulting inclusion probabilities. The resulting sum is denoted n_b ($> n_a$), which may not be an integer value. Next, we adjust again the probabilities for noncertainty units including the units for which the minimum probabilities were imposed as:

$$\pi_i = (n_c - C_q)\left[\frac{\pi_i}{\sum_i^{M_q} \pi_i}\right] \qquad (5)$$

where n_c is the smallest integer value larger than n_b (e.g., if $n_b = 15.3$, then $n_c = 16$). Finally, we add up the resulting (certainty and noncertainty) probabilities. The sum of all these probabilities is the final survey sample size (i.e., the number of surveys to be sent out to), and is denoted n_m ($= n_c$).

Part II: Estimation of Population Parameters and Confidence Intervals

Step 3: Implementation of Pareto Sampling

After the mailout sample size (n_m) for each sector is determined in Step 2, the mailout sample is selected from each sector's population using Pareto sampling. The probability of each unit (vessel) being in the sample in a given sector is proportional to the unit's (vessel's) ex-vessel revenue. Because the majority of gross revenue within each sector comes from a small number of vessels, a random sample of vessels would only include a small portion of the total ex-vessel values.

According to Brewer and Hanif (1983), there are fifty different approaches that are used for UPS. Most of these approaches suffer from the weakness that it is very hard to estimate the variance. Poisson sampling overcomes this problem, and is relatively easy to implement. However, the limitation of Poisson sampling is that the sample size is a random variable. Therefore, in this project, we will use Pareto sampling (Rosen 1997 and Saavedra 1995) which overcomes the limitation of Poisson sampling. The mailout sample size will be n_m as determined in Step 2 (C) above. We will use the inclusion probabilities obtained from Equation (5) above in implementing Pareto sampling.

The procedure of this sampling method (Block and Crowe 2001) is briefly described here:

1. Determine the probability of selection (π_i) for each unit i as in Equation (5) above.
2. Generate a Uniform (0,1) random variable U_i for each unit i
3. Calculate $Q_i = U_i (1 - \pi_i) / [\pi_i (1 - U_i)]$
4. Sort units in ascending order by Q_i, and select n_m smallest ones in sample.

From the above, it is clear that we will have a fixed sample size with Pareto sampling.

Step 4: Mailing out Surveys and Obtaining Actual Response Sample

Next, we will send out the surveys to the n_m units (vessel owners). Actual response sample will be obtained and the size of the actual response sample is denoted r.

Step 5: Estimation of Population Parameters (Population Total)

Using the information in the actual response sample, we calculate population parameters *for variables of interest* (employment and labor income in our project), *not for ex-vessel revenue,* using HT estimator (Horvitz and Thompson 1952). We are interested in estimating the population totals (not population means) of the variables of interest. The HT estimator is given as:

$$\hat{Y}_{HT} = \sum_{i=1}^{r} w_i y_i \qquad (6)$$

where r : number of respondents

w_i : weight for ith unit ($= 1/\pi_i$). Note that the weights are calculated here using the information on the auxiliary variable, not that on the variables of interest

y_i : response sample data of i^{th} unit (employment or labor income)

However, the HT estimator needs to be adjusted for non-response. The estimator is adjusted in the following way.

$$\hat{Y} = \left(\frac{\sum_{j=1}^{N} X_j}{\sum_{i=1}^{r} w_i X_i} \right) \hat{Y}_{HT} \tag{7}$$

where N : population size

X_i : auxiliary variable of i^{th} unit (respondents only)

Usually, we apply this adjustment to the certainties separately from the noncertainties, and then add the two together to get a final estimate. If there are no respondents within any of the two groups of certainty units and noncertainty units, then we collapse the two groups before applying the adjustment. Specifically, the final estimate of population total is given by:

$$\hat{Y} = \left(\frac{\sum_{j=1}^{N_1} X_j}{\sum_{i=1}^{r_1} w_i X_i} \right) \sum_{i=1}^{r_1} w_i y_i \; + \; \left(\frac{\sum_{j=1}^{N_2} X_j}{\sum_{i=1}^{r_2} w_i X_i} \right) \sum_{i=1}^{r_2} w_i y_i \tag{8}$$

where N_1 : number of certainty units in the population

N_2 : number of noncertainty units in the population

r_1 : number of respondents from certainty units

r_2 : number of respondents from noncertainty units, and

$N_1 + N_2 = N$ and $r_1 + r_2 = r$.

Step 6: Estimation of Variance for $\hat{Y}_{HT}$ and $\hat{Y}$

Here we will calculate the variances of the population estimates for the variables of interest. The variance estimate for Pareto sampling is given in Rosen (1997, Equation (4-11), p. 173) as:

$$Var(\hat{Y}_{HT}) = \frac{n_m}{n_m - 1} \left\{ \left[\sum_{i=1}^{n_m} (1-\pi_i) \left(\frac{y_i}{\pi_i} \right)^2 \right] - \frac{\left[\sum_{i=1}^{n_m} y_i \left(\frac{1-\pi_i}{\pi_i} \right) \right]^2}{\sum_{i=1}^{n_m} (1-\pi_i)} \right\} \tag{9}$$

Since we have adjusted for nonresponse, we need to incorporate the variability due to nonresponse into the variance. If we assume that the response mechanism is fixed [8], then we have a ratio estimator and its variance can be found in Hansen, Hurwitz, and Madow (1953, page 514). This variance is a Taylor expansion, and is given as:

$$Var(\hat{Y}) = \hat{Y}^2\left(\frac{\hat{\sigma}^2(A)}{A^2} + \frac{\hat{\sigma}^2(B)}{B^2} - \frac{2COV(A,B)}{AB}\right) \tag{10}$$

where

$$A = \sum_{i=1}^{r} w_i y_i$$

$$B = \sum_{i=1}^{r} w_i X_i$$

$$\hat{\sigma}^2(A) = \frac{n_m}{n_m - 1}\left\{\left[\sum_{i=1}^{r}(1-\pi_i)(w_i y_i)^2\right] - \frac{\left[\sum_{i=1}^{r}(1-\pi_i)(w_i y_i)\right]^2}{\sum_{i=1}^{n_m}(1-\pi_i)}\right\}$$

$$\hat{\sigma}^2(B) = \frac{n_m}{n_m - 1}\left\{\left[\sum_{i=1}^{r}(1-\pi_i)(w_i X_i)^2\right] - \frac{\left[\sum_{i=1}^{r}(1-\pi_i)(w_i X_i)\right]^2}{\sum_{i=1}^{n_m}(1-\pi_i)}\right\}$$

$$COV(A,B) = \frac{n_m}{n_m - 1}\left\{\left[\sum_{i=1}^{r}(1-\pi_i)w_i^2 y_i X_i\right] - \frac{\left[\sum_{i=1}^{r}(1-\pi_i)(w_i y_i)\right]\left[\sum_{i=1}^{r}(1-\pi_i)(w_i X_i)\right]}{\sum_{i=1}^{n_m}(1-\pi_i)}\right\}.$$

<u>Step 7: Calculation of Confidence Intervals</u>

Confidence intervals are calculated using response sample statistics obtained in steps 5 and 6. We only choose one sample, but if there were many independent samples chosen then we would expect on average that approximately 100(1-α) % of the confidence intervals constructed in the following manner will contain the truth.

$$\left(\hat{Y} - z_{\alpha/2}\sqrt{Var(\hat{Y})}\ ,\ \hat{Y} + z_{\alpha/2}\sqrt{Var(\hat{Y})}\right) \tag{11}$$

where $\hat{Y}$: Estimated population total for employment or labor income.

Note that it is possible to use t-statistics if the sample size is small.