An Estimate for the Biased Sample and the Non-Bernoulli Sampling Process

J. Goodman
Bechtel Power Corporation, 12400 E. Imperial Highway, Norwalk, California 90650, U.S.A.

Abstract

Classical statistical techniques for inspection-by-attributes sampling plans are based on assumptions of a random sample and Bernoulli sampling process. However, for many practical cases of sampling inspections of nuclear power plants, these assumptions are not valid. A statistical estimation technique using weaker assumptions is developed. This technique uses cluster sampling in conjunction with Monte Carlo simulation to derive a standard and the best confidence interval for the frequency of defects.

1. Introduction

Sampling by attributes is based on two assumptions:

(1) The sample is random and unbiased, i.e., each item in the population has an equal chance of being selected;

(2) Any randomly selected item has an equal chance of being defective (so-called Bernoulli sampling process).

For many practical cases of sampling inspections of nuclear power plants, the preceding assumptions are not valid. Because of the existence of a large uncountable population or inaccessability of part of the population, it is not possible to draw a random unbiased sample. Also, because work is done by different worker teams, at different periods of time, or different suppliers are utilized, the rate of defects can be different for different portions of the population. Nevertheless, we need reliable estimates for these cases, although classical statistics does not address them. In this paper we develop a method to estimate the proportion of defects in the population which does not rely on classical statistical assumptions.

2. Cluster Sampling

Assume that the population is divided into N clusters. If the subpopulation size of the ith cluster is designated as n_i and frequency of defects in the subpopulation as p_i, then the frequency of defects in the population is:

$$P = \frac{1}{n_N} \sum_{i=1}^{N} n_i \, p_i \qquad (1)$$

where n_N is a total population size:

$$n_N = \sum_{i=1}^{N} n_i \qquad (2)$$

Traditional procedures of cluster sampling require a random sampling of ν out of N clusters, and then random independent sampling (or a complete census) in every randomly selected cluster. The sample estimate for frequency $\hat{P}$ and variance $V(\hat{P})$ according to Cochran [1] is:

$$\hat{P} = \frac{1}{n_\nu} \sum_{j=1}^{\nu} n_j \, p_j \qquad (3)$$

$$V(\hat{P}) = \frac{N(N-\nu)}{\nu} \cdot \frac{\sum_{j=1}^{N} n_j^2 \, (p_j - P)^2}{N-1} \qquad (4)$$

where

$$n_\nu = \sum_{j=1}^{\nu} n_j \qquad (5)$$

To assess the variance (4) and, therefore, the confidence interval for the frequency of defects (1) we need to know the size of all subpopulations n_i. Different approximations (see, for example, Satterthwaite [2]) use the assumption that the frequency of defects in all subpopulations is essentially the same, with only small fluctuations from cluster to cluster which can be described with the normal distribution. Another sensitive point of these approximations is determination of the effective number of degrees of freedom. Therefore, a regular cluster sample procedure cannot solve our problem. However, we may apply a modified cluster approach based on different assumptions.

3. New Assumptions

In our analysis two assumptions were adopted:

(1) The cluster sizes n_i for several randomly selected accessible clusters can be determined;

(2) The set of numbers n_i and p_i for the clusters in the sample population can be fitted with a smooth distribution f(n, p).

Due to the first assumption, the problem of accountability or assessability eases because information is only required about a few accessible clusters rather than the entire population.

The second assumption practically poses no limitation if the clusters are delimited carefully. Let n and p be variables which take on values of n_i and p_i. If we take a range for n or p and divide it into several intervals we will create a histogram of the distribution. There are several possibilities: the histogram may have a maximum or a minimum, be monotonely increasing or decreasing. In these cases it is obvious how to fit the histogram with a smooth function. A more difficult case is when the histogram has an erratic form. In this case, probably the best solution is to fit the histogram with

a uniform distribution reflecting all fluctuations in uncertainty of parameters. In the same manner, the problem can be solved for the joint distribution of n and p if the correlation matrix indicates a significant correlation between the two distributions.

4. Evaluation of Cluster Sizes and Frequencies of Defects for Selected Clusters

The size of a cluster should be selected so that it is feasible to count the number of items in the subpopulation n_i by a walkthrough, by an examination of drawings and documents, or by approximation. Another suggested requirement is that, within a cluster, the subpopulation be as uniform as possible. Uniformity within clusters reduces the total variance. In the case of accessibility restrictions, we have an opportunity to make an evaluation based on a biased sample using the likelihood density function method (see Goodman [3], [4]).

If it is impossible to count exactly the number n_i, then its estimated value can be used. The assigned value should include both a best estimate and uncertainty. The lognormal distribution of uncertainty can then be applied according to the principle of the maximum entropy (see Goodman [4]).

If 100 percent inspection of selected clusters is used, then frequencies p_i are known exactly; otherwise a random subsampling procedure can be used. The distribution of uncertainty of p_i can be described using likelihood density function method. The advantage of this method is that it does not require a random unbiased sample. The only restrictions are: (1) The sample must be taken within the same population; (2) All measurements must be independent (i.e., the result of current measurement is not affected by results of previous ones).

5. Evaluation Cluster Sizes and Frequencies of Defects for Uninspected Clusters

Let n_i and p_i be exact values or estimated values for several selected clusters. To determine the n_i and p_i for other uninspected clusters, we will use the likelihood density function method.

We begin by defining the analytical form of the joint distribution $f(n, p; \alpha_1, \alpha_2, \ldots \alpha_m)$ where $\alpha_1, \alpha_2, \ldots \alpha_m$ are parameters of the distribution. The principle of maximum entropy mentioned above is a useful tool at this point.

The next step is a contribution of the likelihood function L according to the following formula:

$$L(\alpha_1, \alpha_2, \ldots \alpha_m) = \prod_{i=1}^{\nu} f(n_i, p_i; \alpha_1, \alpha_2, \ldots \alpha_m) \qquad (6)$$

The likelihood density function $\phi(\alpha_1, \alpha_2, \ldots \alpha_m)$ is determined as:

$$\phi(\alpha_1, \alpha_2, \ldots \alpha_m) = \frac{L(\alpha 1, \alpha 2, \ldots \alpha m)}{\int\int \ldots \int L(\alpha_1, \alpha_2, \ldots \alpha_m) d\alpha_1, d\alpha_2 \ldots d\alpha_m} \qquad (7)$$

Therefore, the probability of dP(n, p) that a randomly selected cluster will have a size within an interval (n, n+dn) and frequency of defects within a range (p, p+dp) is

$$dP(n, p) = f(n, p; \alpha_1, \alpha_2, \ldots \alpha_m)\, dndp \qquad (8)$$

where random parameters α_1, α_2, ... α_m are distributed accordingly to the joint distribution $\phi(\alpha_1, \alpha_2, \ldots \alpha_m)$.

6. Evaluation of the Frequency of Defects in the Population

For a point estimate of the frequency of defects in the population, the formula (1) is used in conjunction with three-level Monte Carlo simulation. At the first step we randomly select those n_i and p_i for ν selected clusters which have some uncertainty. If all n_i and p_i are known exactly (countable subpopulation of clusters and 100 percent inspection of every selected cluster), then the first step should be omitted.

At the second step, we randomly draw the set of random parameters $\alpha_1, \alpha_2, \ldots \alpha_m$ according to the distribution (7) which can be constructed after the first step.

At the third step, we randomly generate n_i and p_i for the remaining unexamined clusters using the distribution function (8). As a result of three simulations we have a set of random n_i and p_i which are used for evaluating the frequency of defects P according to the formula (1).

Repeating this procedure many times we generate the distribution of the frequency P, and then can determine median, lower and upper limits, mean and variance, and the best confidence interval (see Goodman [5]).

7. Conclusion

The proposed technique can be applied using less restrictive assumptions than regular sampling procedures. It can be used when sampling is unavoidably non-random or biased, and even when the underlying process is non-Bernoulli. The flexibility of the method makes it optimally suited for sampling applications in the field where it is frequently not possible to satisfy the statistical assumptions about a random sample and Bernoulli sampling process.

In this paper sampling by attributes was considered. A similar technique can be used for a sampling by variables. The details will be discussed elsewhere.

References

[1] COCHRAN, W. G., Sampling Technique, New York: John Wiley and Sons, 1977.

[2] SATTERTHWAITE, F. E., "An Approximate Distribution of Estimates of Variance Components," Biometrics, 2, 110-114, 1946.

[3] GOODMAN, J., "Classification and Comparison of Sampling Procedures," Transactions of the 8th International Conference on Structural Mechanics in Reactor Technology, Paper M2 2/8, Brussels, Belgium, August 19-23, 1985.

[4] GOODMAN, J., "Estimating Fragility Curves Using Few Experimental Data," Probabilistic Structural Analysis, PVP-Vol. 93, ASME, 41-61, 1984.

[5] GOODMAN, J., "On the Definition of the "Best" Confidence Interval," Reliability Engineering, 7, 213-228, 1984.