Use of Two-Stage or Double Sampling in Final Status Decommissioning Surveys

Carl V. Gogolak

When might it be desirable to allow a licensee to sample a survey unit a second time to determine compliance? In the Statistical literature this is called either two-stage sampling or double sampling. Resampling is something else altogether. The terms "double sampling" and "two-stage sampling" seem to appear interchangeably in the literature. More recently, the later seems to have gained favor, so we will use two-stage sampling in this paper when referring to survey designs specifically intended to be conducted in two stages. We will use the term double sampling to refer to the case when the survey design is a one stage design, but allowance is made for a second set of samples to be taken if the retrospective power of the test using the first set of samples does not meet the design objectives. Such allowance, if given, should be specifically mentioned in preparing the DQOs and in advance of any sampling and analysis. During the DQO process, double sampling could be considered as an option in setting the Type I error rates. The reasoning behind this is discussed in the next section.

Double Sampling

Suppose it is thought that a survey unit might have passed the final status statistical test had the initial sampling design been powerful enough. That is, a retrospective examination of the power of the statistical tests used reveals that the probability of detecting that the survey unit actually meets the release criterion was lower than that planned for during the DQO process. This could occur if the spatial variability in residual radioactivity concentrations was larger than anticipated. The power of the test specified during the DQO process depends on an estimate of the uncertainty. The power of the statistical test will be less than planned if the standard deviation is higher than expected. If samples were lost, did not pass analytical QA/QC, or are otherwise unavailable for inclusion in the analysis, the power will also be lower than was planned. Might additional samples be taken in the survey unit to improve the power of the test?

The Draft NUREG/CR-5849 allowed the licensee to take additional samples in a survey unit if, after the first sampling, the mean was less than the DCGL, but the desired upper confidence level on the mean was not. Because a 95% confidence interval is constructed using Student's t statistic rather than using a hypothesis test, Type II errors are not considered in the survey design. The second set of samples was taken so that a t test on the combined set of samples would have 90% power at the mean of the first set of samples, given the estimated standard deviation from the first set of samples. Such double sampling was to be allowed only once.

Increasing the probability that a clean a survey unit passes (power) by the use of Double Sampling will also tend to increase the probability that a survey unit that is not clean will pass (type I error). In

addition, the two tests are not independent because the data from the first set of samples is used in both. The increase in the Type I error rate is probably less than a factor of two. But the fact that this is possible when Double Sampling is allowed should be clearly understood at the beginning. Thus, the issue of whether or not to allow Double Sampling is properly a part of the DQO process used to set the acceptable error rates.

Two-stage or double sampling is not usually expected (nor is it encouraged) when the DQO process is used, as in the MARSSIM. This is because the Type II error and the power desired are explicitly considered in the survey design process. If higher power in the test is desired, it should be specified as such. Sufficient samples should be taken to achieve the specified power. The value of this approach lies in the greater objectivity and defensibility of the decision made using the data. Nonetheless, it is recognized that there may be instances when some sort of double sampling is considered desirable. For example, when it is difficult to estimate the standard deviation of the concentrations in a survey unit. A first set of data may be taken with an estimated standard deviation that is too low, and thus, the power specified in the DQO process may not be achieved. Similarly, some pilot data may be taken to estimate the standard deviation in a survey unit. Under what circumstances may this data also be used in the test of the final status?

In such cases, it will be useful for planning if there is an estimate of how much the type II error rate might increase as a result of double sampling.

Consider the Sign test. As indicated in NUREG-1505 Rev.1. Suppose N_1 samples are taken. Recall that for the Sign test in Scenario A, the test statistic, S_1, was equal to the number of survey unit measurements below the $DCGL_W$. If S_1 exceeds the critical value k_1, then the null hypothesis that the median concentration in the survey unit exceeds the $DCGL_W$ is rejected, i.e., the survey unit passes this test. The probability that any single survey unit measurement falls below the $DCGL_W$ is found from

$$p(C) = \int_{-\infty}^{DCGL_W} f(x)dx = \frac{1}{\sqrt{2p}\,s} \int_{-\infty}^{DCGL_W} e^{-(x-C)^2/2s^2} dx = \Phi\left(\frac{DCGL_W - C}{s}\right)$$

C is the true, but unknown, mean concentration in the survey unit. When $C = DCGL_W$, $p = 0.5$.

The probability that more than k_1 of the N_1 survey unit measurements fall below the $DCGL_W$ is simply the following binomial probability:

$$\sum_{t=k_1+1}^{N_1} \binom{N_1}{t} p^t (1-p)^{N_1-t} = 1 - \sum_{t=0}^{k_1} \binom{N_1}{t} p^t (1-p)^{N_1-t}$$

This is the probability that the null hypothesis will be rejected, and it will be concluded that the survey unit meets the release criterion. When the mean concentration in the survey unit is at the $DCGL_W$, this is just the Type I error rate, α.. When $C = DCGL_W$, $p = (1-p) = 0.5$, *so*

$$a = \sum_{t=k_1+1}^{N_1} \binom{N_1}{t} (0.5)^t (0.5)^{N_1 - t} = (0.5)^{N_1} \sum_{t=k_1+1}^{N_1} \binom{N_1}{t}$$

Now, suppose it is decided to allow the licensee to take a second set of samples of size N_2. The test statistic, S, is equal to the number of the total of $N = N_1 + N_2$ survey unit measurements below the $DCGL_W$. If S exceeds the critical value k , then the null hypothesis that the median concentration in the survey unit exceeds the $DCGL_W$ is rejected, i.e., the survey unit passes this test. Now the overall probability that the null hypothesis is rejected (i.e., the survey unit passes) is equal to the sum of the probabilities of two events that are mutually exclusive:

1) The probability that more than k_1 of the N_1 survey unit measurements fall below the $DCGL_W$

and

2) The probability that fewer than k_1 of the *first* N_1 survey unit measurements fall below the $DCGL_W$ but that more than k of the N total survey unit measurements fall below the $DCGL_W$.

Now $S = S_1 + S_2$, where S_2 is the number of the second set of N_2 survey unit measurements that fall below the $DCGL_W$. S_1 and S_2 are independent, but S_1 and $S = S_1 + S_2$ are not.
The covariance of S_1 and S , using $E(\cdot)$ to denote expected value, is

$$\begin{aligned} Cov(S_1, S) &= E(S_1 S) - E(S_1)\,E(S) \\ &= E(S_1(S_1 + S_2) - E(S_1)\,E(S) \\ &= E(S_1^2) + E(S_1 S_2) - E(S_1)\,E(S) \\ &= (N_1^2 p(1-p) + N_1^2 p^2) + N_1 N_2 p^2 - N_1 p(N_1 + N_2) p \\ &= N_1 p(1-p) \end{aligned}$$

Therefore the correlation coefficient between S_1 and S is

$$
\begin{aligned}
r(S_1, S) &= \frac{N_1 p(1-p)}{\sqrt{N_1 p(1-p)(N_1+N_2)\,p(1-p)}} \\
&= \frac{N_1}{\sqrt{N_1(N_1+N_2)}} \\
&= \sqrt{N_1/(N_1+N_2)} = \sqrt{N_1/N}
\end{aligned}
$$

To calculate the overall probability that the survey unit passes, we require the joint probability of S_1 and S,

$$
\begin{aligned}
\Pr(S_1 = s_1, S = s) &= \Pr(S_1 = s_1)\Pr(S_2 = s - s_1) \\
&= \binom{N_1}{s_1} p^{s_1}(1-p)^{N_1 - s_1} \binom{N_2}{s - s_1} p^{s - s_1}(1-p)^{N_2 - (s - s_1)} \\
&= \binom{N_1}{s_1}\binom{N_2}{s - s_1} p^{s}(1-p)^{N - s}
\end{aligned}
$$

Therefore, the overall probability that the survey unit passes is

$$
\begin{aligned}
\Pr(S_1 > k_1 \text{ or } S > k) &= \Pr(S_1 > k_1) + \Pr(S_1 \le k_1 \text{ and } S > k) \\
&= \sum_{s_1 = k_1 + 1}^{N_1} \binom{N_1}{s_1} p^{s_1}(1-p)^{N_1 - s_1} \\
&\quad + \sum_{s_1 \le k_1} \sum_{s_2 > k - s_1} \binom{N_1}{s_1}\binom{N_2}{s_2} p^{s_1 + s_2}(1-p)^{(N_1 + N_2) - (s_1 + s_2)}
\end{aligned}
$$

The first term is equal to (or slightly less than) the Type I error rate α specified during the DQO process. The second term is the additional probability of a Type I error introduced by allowing double

sampling.

Note that

$$\Pr(S_1 \le k_1 \text{ and } S > k) = \sum_{s>k}^{N} p^s (1-p)^{N-s} \sum_{s_1=0}^{k_1} \binom{N_1}{s_1}\binom{N_2}{k-s_1}$$

$$\le \sum_{s>k}^{N} p^s (1-p)^{N-s} \sum_{s_1=0}^{k} \binom{N_1}{s_1}\binom{N_2}{k-s_1}$$

$$= \sum_{s>k}^{N} p^s (1-p)^{N-s} \binom{N}{s} = \Pr(S > k) \le a$$

Thus the Type I error rate would be at most doubled when double sampling is allowed.

For example, if a survey is designed so that N_1 = 30, and α = 0.05. Then the critical value for the Sign test is k_1 = 19. Suppose the first survey results in 19 or fewer measurements less than the $DCGL_W$. In addition, suppose the survey unit is sampled again, taking an additional N_2 = 30 samples. Then the total number of samples is $N = N_1 + N_2 = 60$. The critical value for the Sign test with α = 0.05 and N = 60 is k = 36. When the survey unit concentration is equal to the $DCGL_W$, $p = 0.5$, so we have

$$\Pr(S_1 > 19 \text{ or } S > 36) = \Pr(S_1 > 19) + \Pr(S_1 \le 19 \text{ and } S > 36)$$

$$= \sum_{s_1=20}^{30} \binom{30}{s_1} (0.5)^{s_1} (1-0.5)^{30-s_1}$$

$$+ \sum_{s_1=0}^{19} \binom{30}{s_1} \sum_{s_2=(37-s_1)}^{30} \binom{30}{s_2} (0.5)^{s_1+s_2} (1-0.5)^{(30+30)-(s_1+s_2)}$$

$$= 0.049 + 0.027 = 0.076$$

Thus the total Type I error rate is about 50% greater than originally specified.

In conclusion, double sampling should not be used as a substitute for adequate planning. If it is to be allowed, this should be agreed upon as part of the DQO process. The procedure for double sampling, i.e. the size of the second set of samples, N_2 , should be specified, recognizing that the Type I error rate could be up to twice that specified for the Sign test when only one set of samples is taken.

Similar consideration apply for the WRS test, however the calculation of the exact effect on the Type I

error rate is considerably more complex.

Finally, we note that double sampling should never be necessary for Class 2 or Class 3 surveys, which are not expected to have concentrations above the $DCGL_W$. These classes of survey unit should always pass after the first set of samples because every measurement should be below the $DCGL_W$. The very need for a second set of samples (i.e. failure to reject the null hypothesis) in Class 2 or Class 3 survey units would raise an issue of survey unit mis-classification. In addition, double sampling is generally not appropriate for Class 1 Survey units where elevated areas have been found.

A better solution to the issue of Double Sampling is to plan for data collection in two stages, and design the final status survey accordingly, as is discussed in the remainder of this report.

Two-Stage Sequential Sampling

Suppose there are a large number of survey units of a similar type to be tested. In this case a two-stage sampling procedure may result in substantial savings by reducing the average number of samples required to achieve a given level of statistical power.

Suppose we desire to plan a two-stage sign test. Let N_1 be the size of the first set of samples taken, and let $\boldsymbol{S}_1$ be the number of these less than the DCGL. Similarly, let N_2 be the size of the second set of samples taken, and let $\boldsymbol{S}_2$ be the number of these less than the DCGL. Let $N = N_1 + N_2$, and let $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$. The procedure is as follows:

if $\boldsymbol{S}_1 > u_1$ then the survey unit passes (reject H_0)
if $\boldsymbol{S}_1 < l_1$ then the survey unit fails
if $l_1 \le \boldsymbol{S}_1 \le u_1$ then the second set of samples is taken.

If $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2 > u_2$ after the second set of samples is analyzed, then the survey unit passes.

What is the advantage of two-stage testing? For given error rates α and β, the number of samples, N_1, taken in the survey unit during the first stage of sampling will be less than the number, N_0, required in the MARSSIM tables. Unless the result is "too close to call", this will be the only sampling needed. When the result is "too close to call", $l_1 \le \boldsymbol{S}_1 \le u_1$, a second sample of size N_2 is taken and the test statistic $\boldsymbol{S}_2$ is computed using the combined data set, $N_1 + N_2$. While the size of the combined set, $N = N_1 + N_2$, will generally be larger than the number, N_0, from in the MARSSIM tables, the expected sample size over many survey units is still lower. Thus two-stage sampling scheme will be especially useful when there are many similar survey units for which the final status survey design is essentially the same. Two-stage sampling may be used whether or not a reference area is needed. It may be used with either the Sign or the WRS test.

Now, the major issue is how to choose the critical values l_1 , u_1 , and u_2 . Hewitt and Spurrier (1983)

suggest three criteria:

1) Match the power curve of the two-stage test to that of the one-stage test. The curves are matched at three points. The points with power equal to α, $1 - \beta$, and 0.5 are generally well enough separated to assure a good match over the entire range of potential survey unit concentrations.

2) Maximize the power at the LBGR for given values of α and average sample size.

3) Minimize the sample size for given values of α, and $1-\beta$.

While any one of these criteria could be used, the first has received more attention in the literature. Thus, it may be more readily applied to the case of final status survey design. The other criteria would require further development.

Spurrier and Hewitt (1975) initially developed a two stage sampling methodology using criteria 1 assuming the data are normally distributed. They matched power at α, 0.5 and 0.9. Table 1 shows the values of l_1 , u_1 , and u_2 they obtained for six different sets of sample sizes, N_1/N_0 , N_2/N_0 , expressed as fractions of the sample size, N_0 , that would be required for the one stage test with equivalent power. The term $E(N)/N_0$, is the maximum expected combined sample size for the two stage test relative to the sample size, N_0 , that would be required for the one stage test with equivalent power. This number is almost always less than one, but it depends on how close the actual concentration in the survey unit is to the $DCGL_W$. Clearly, if the concentration is over the $DCGL_W$, the survey unit is likely to fail on the first set of samples. If the concentration is much lower than the $DCGL_W$, the survey unit is likely to pass on the first set of samples. It is only when the true concentration in the survey unit falls within the gray region that there will be much need for the second set of samples. The fact that the maximum $E(N)/N_0$ is almost always less than one indicates that overall number of samples required for a two stage final status survey will almost never exceed the number required for a one stage test, even if the true concentration the survey unit falls in the gray region between the LBGR and the $DCGL_W$.

Recall that the power to distinguish clean from dirty survey units is relativley low when the true concentration is in the gray region. It falls from 1-β at the LBGR to α at the $DCGL_W$. Thus ,when the true concentration is in the gray region, there will be a larger of cases when the second set of samples is needed. The gray region is exactly where the results are "too close to call". However, if the true concentration the survey unit is below the LBGR or above the $DCGL_W$, the actual average number of samples will be closer to N_1, because the second set of samples will seldom be needed.

In 1976, Spurrier and Hewitt dropped the assumption of normality and extended their methodology to two-stage Wilcoxon Signed Rank (WSR) and Wilcoxon Rank Sum (WRS) tests. The procedure depends on an extension of the Central Limit Theorem to the joint distribution of the test statistics $\boldsymbol{S}_1$ and $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$. Spurrier and Hewitt suggest that the approximation works reasonably well for sample

sizes as small as 9.

In this paper, we will apply their method to the Sign test as well.

For the Sign test, we compute

$$S_1 = \frac{S_1^+ - N_1/2}{\sqrt{N_1/4}},$$

where S_1^+ is the usual Sign Test statistic, i.e. the number of measurements less than the $DCGL_W$.

Using Table 1,

if $\boldsymbol{S}_1 > u_1$ then reject the null hypothesis (the survey unit passes)
if $\boldsymbol{S}_1 < l_1$ then do not reject the null hypothesis (the survey unit fails)
if $l_1 \le \boldsymbol{S}_1 \le u_1$ then take the second set of samples.

If a second set of samples is taken, then compute

$$S = \frac{(S_1^+ + S_2^+) - (N_1 + N_2)/2}{\sqrt{(N_1 + N_2)/4}} = \frac{S^+ - N/2}{\sqrt{N/4}}$$

Using Table 1,

if $\boldsymbol{S} > u_2$ then reject the null hypothesis (the survey unit passes)
if $\boldsymbol{S} \le u_2$ then do not reject the null hypothesis (the survey unit fails).

This test relies on "a large sample approximation". That is, we are assuming that the sample size is large enough that the joint distribution of $\boldsymbol{S}_1$ and $\boldsymbol{S}$ is bivariate standard normal with correlation coefficient $r(S_1, S) = \sqrt{N_1 / N}$. Some simulation studies would be needed to determine quantitative bounds on the accuracy of this approximation.

The choice of which set of sample sizes should be used is dependent on how confident one is of passing.

For Class 2 and Class 3 survey units, case 3 with $N_1/N_0 = 0.2$ and $N_2/N_0 = 1.0$ might be reasonable. In these Classes of survey units no individual sample concentrations in excess of the $DCGL_W$ are expected. The probability of passing on the first set of samples should be close to one. Therefore, it makes sense to choose a design with the minimum number of samples required in the first set.

For Class 1 survey units, case 2 with $N_1/N_0 = 0.4$ and $N_2/N_0 = 0.8$ might be more appropriate. There is some chance that the survey unit will not pass on the first set of samples, so it may be desirable to

reduce Max $E(N)/N_0$ from 0.999 to 0.907 by taking more samples in the first set.

If the gray region has been expanded in order to increase Δ/σ, case 1 or 4 would be a more conservative choice. In this situation, statistical power has been compromised somewhat, so it may be important to reduce the risk of having a larger average total number of samples (as indicated by the potential Max $E(N)/N_0$) even further.

Scan sensitivity will also impact the ability to use two stage designs in Class 1 survey units. It would have to be determined if the $DCGL_{EMC}$ can be detected when only N_1 samples are taken. If not, the sample size would have to be increased until the scan MDC is lower than the $DCGL_{EMC}$. In this situation, the choice of N_1, and the average savings possible with two-stage sampling may be severely limited.

Table 1 Critical Points for Two Stage Test of Normal Mean for a One Sided Alternative
Source: Spurrier and Hewett (1975).

			$\alpha = 0.05$				$\alpha = 0.01$			
	N_1/N_0	N_2/N_0	u_1	l_1	u_2	Max $E(N)/N_0$	u_1	l_1	u_2	Max $E(N)/N_0$
1	0.60	0.60	1.886	0.710	1.783	0.866	2.499	1.259	2.493	0.879
2	0.40	0.80	1.984	0.179	1.782	0.907	2.558	0.635	2.496	0.931
3	0.20	1.00	2.073	-0.482	1.784	0.999	2.600	-0.146	2.502	1.030
4	0.55	0.55	2.050	0.438	1.716	0.869	2.635	0.966	2.411	0.878
5	2/3	2/3	1.781	0.950	1.868	0.882	2.415	1.520	2.600	0.897
6	0.70	0.70	1.749	1.045	1.909	0.893	2.390	0.628	2.651	0.908

For the WRS test, at each stage we set the number of measurements required in the survey unit, n_1 and n_2 , and in the reference area m_1 and m_2 relative to the number required for the one stage test $n_0 = m_0 = N_0/2$ specified in Table 5.3 of the MARSSIM. There is an additional requirement that $n_1/n_2 = m_1/m_2$, which should be satisfied with sufficient accuracy for most MARSSIM designs. Minor departures due to small differences in sample size caused by filling out systematic grids or the loss of a few samples should not severely impact the results.

We compute

$$S_1 = \frac{W_1^R - m_1(n_1 + m_1 + 1)/2}{\sqrt{n_1 m_1 (n_1 + m_1 + 1)/12}},$$

where W_1^R is the usual WRS Test statistic, i.e. the sum of the ranks of the adjusted reference area measurements.

Using Table 1,

if $\boldsymbol{S}_1 > u_1$ then reject the null hypothesis (the survey unit passes)
if $\boldsymbol{S}_1 < l_1$ then do not reject the null hypothesis (the survey unit fails)
if $l_1 \le \boldsymbol{S}_1 \le u_1$ then take the second set of samples.

If a second set of samples is taken, then compute

$$S = \frac{(W_1^+ + W_2^+) - (m_1 + m_2)(m_1 + m_2 + n_1 + n_2 + 1)/2}{\sqrt{(m_1 + m_2)(n_1 + n_2)(m_1 + m_2 + n_1 + n_2 + 1)/12}} = \frac{W^R - m(m + n + 1)/2}{\sqrt{mn(m + n + 1)/12}}$$

Using Table 1,

if $\boldsymbol{S} > u_2$ then reject the null hypothesis (the survey unit passes)
if $\boldsymbol{S} \le u_2$ then do not reject the null hypothesis (the survey unit fails).

This test relies on "a large sample approximation". That is, we are assuming that the sample size is large enough that the joint distribution of $\boldsymbol{S}_1$ and $\boldsymbol{S}$ is bivariate standard normal with correlation coefficient

$$r(S_1, S) = \sqrt{(m_1 + n_1)/(m + n)}.$$

Some simulation studies would be needed to determine some quantitative bounds on the accuracy of this approximation.

An Alternative Two-Stage Two-Sample Median Test

A different approach to this testing problem has been suggested by Wolfe (1977). In his procedure, a specific number of sample measurements are made in a reference area, and the median, M, is

calculated, and the $DCGL_W$ added. Survey unit samples are then analyzed until r of them are found to be below M. The test statistic, n_r, is the number of survey unit samples that have been analyzed. Smaller values of n_r indicate that the survey unit meets the release criterion. For Class 2 and Class 3 survey units in particular, we would expect that $n_r = r$. In that case, the number of reference area measurements, m, and the value of r are chosen to meet the DQO for the Type I error rate. In each survey unit, r samples are taken. If all are less than M, we reject the null hypothesis that the survey unit exceeds the release criterion. If any one of them exceeds M, the null hypothesis will not be rejected. Thus, the total number of samples needed in each survey unit may be relatively small. In addition, as soon as one sample is measured above M, the result of the test is known. Thus it may not be necessary to analyze every survey unit sample. Of course, the need to identify elevated areas may preclude the use of this method in some circumstances. However, the potential savings when the analytical costs are high may make this procedure attractive. It merits further investigation.

References

Hewitt, J.E. and J. D. Spurrier, 1983
A Survey of Two Stage Tests of Hypothesis: Theory and Application
Communications on Statistics - Theory and Methods, 12, 20, 2307-2425.

Spurrier, J. D., and J. E. Hewett, 1975
Double Sample Tests for the Mean of a Normal Distribution
Journal of the American Statistical Association, 70, 350, 448-450

Spurrier, J. D., and J. E. Hewett, 1976
Two-Stage Wilcoxon Tests of Hypotheses
Journal of the American Statistical Association, 71, 356, 982-987

Wolfe, D. A., 1977
Two-Stage Two Sample Median Test
Technometrics 19, 4, 495-501