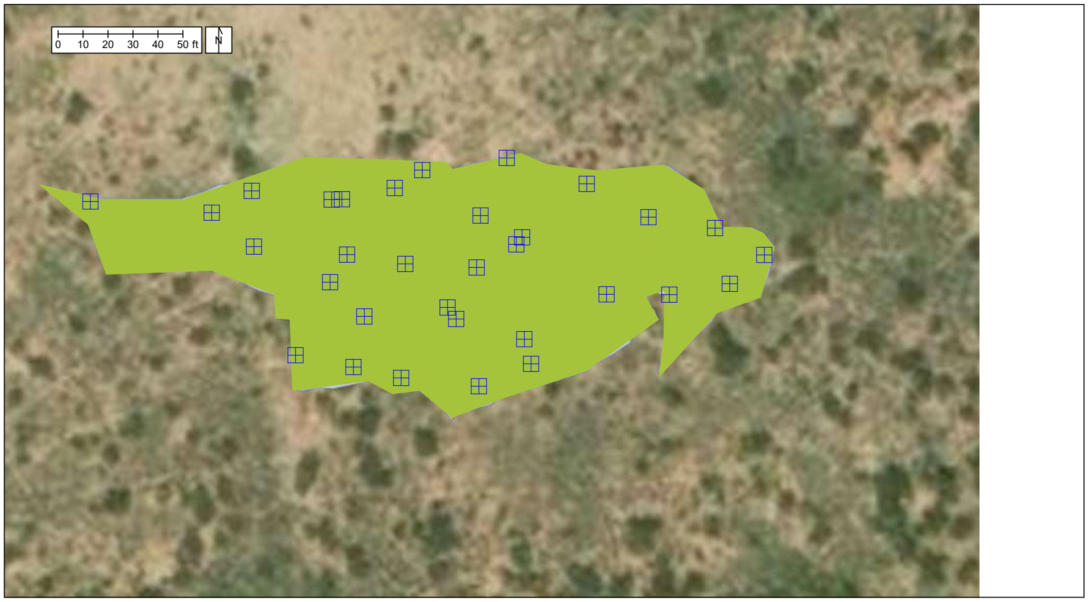
**Sampling Requirements for Finding the Upper Confidence Limit (UCL) on a Proportion**

**Summary**
This report summarizes the sampling design, associated statistical assumptions, as well as general guidelines for conducting post-sampling data analysis. Sampling plan components presented here include how many sampling locations to choose and where within the sampling area to collect those samples. The type of medium to sample (i.e., soil, groundwater, etc.) and how to analyze the samples (in-situ, fixed laboratory, etc.) are addressed in other sections of a sampling plan.

The following table summarizes the sampling design.

| SUMMARY OF SAMPLING DESIGN | |
|---|---|
| Primary Objective of Design | Compute the upper confidence limit on a proportion |
| Sample Placement (Location) in the Field | Random sampling within grids |
| Formula for calculating number of sampling locations | Standard normal approximation of the binomial distribution |
| Confidence level | 90% |
| Maximum difference between estimated proportion and UCL | 0.05 |
| Using prior knowledge about expected proportion | Yes |
| Maximum expected proportion | 0.05 |
| Calculated total number of samples | 32 |
| Number of samples on map [a] | 32 |
| Number of selected sample areas [b] | 1 |
| Specified sampling area [c] | 17802.95 $ft^2$ |
| Total cost of sampling [d] | $17,000.00 |

[a] This number may differ from the calculated number because of 1) grid edge effects, 2) adding judgment samples, or 3) selecting or unselecting sample areas.
[b] The number of selected sample areas is the number of colored areas on the map of the site. These sample areas contain the locations where samples are collected.
[c] The sampling area is the total surface area of the selected colored sample areas on the map of the site.
[d] Including measurement analyses and fixed overhead costs. See the Cost of Sampling section for an explanation of the costs presented here.




| Area: Area 1 | | | | | | |
|---|---|---|---|---|---|---|
| X Coord | Y Coord | Label | Value | Type | Historical | Sample Area |
| -333.1783 | -195.2921 | | | Random in Grid | | |
| -314.1400 | -199.7545 | | | Random in Grid | | |
| -282.9638 | -203.0744 | | | Random in Grid | | |
| -262.1270 | -194.0583 | | | Random in Grid | | |
| -356.4098 | -190.5425 | | | Random in Grid | | |
| -328.8490 | -175.1046 | | | Random in Grid | | |
| -295.5195 | -171.5060 | | | Random in Grid | | |
| -292.0604 | -176.1890 | | | Random in Grid | | |
| -264.7704 | -184.2341 | | | Random in Grid | | |
| -372.9912 | -147.2345 | | | Random in Grid | | |
| -342.4906 | -161.3747 | | | Random in Grid | | |
| -335.7158 | -150.4002 | | | Random in Grid | | |
| -312.4420 | -154.1554 | | | Random in Grid | | |
| -283.9138 | -155.4806 | | | Random in Grid | | |
| -268.0395 | -146.2974 | | | Random in Grid | | |
| -231.9467 | -166.3224 | | | Random in Grid | | |
| -206.7178 | -166.4543 | | | Random in Grid | | |
| -182.5145 | -162.0564 | | | Random in Grid | | |
| -168.7044 | -150.4612 | | | Random in Grid | | |
| -438.4993 | -129.1148 | | | Random in Grid | | |
| -389.9364 | -133.6290 | | | Random in Grid | | |
| -373.9259 | -124.8376 | | | Random in Grid | | |
| -341.8796 | -128.3052 | | | Random in Grid | | |
| -337.7957 | -128.1987 | | | Random in Grid | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| -316.7276 | -123.7170 | | | Random in Grid | | |
| -282.3015 | -134.7397 | | | Random in Grid | | |
| -265.7391 | -143.4883 | | | Random in Grid | | |
| -239.7792 | -122.0766 | | | Random in Grid | | |
| -215.1125 | -135.3722 | | | Random in Grid | | |
| -188.4201 | -139.8008 | | | Random in Grid | | |
| -305.6516 | -116.5408 | | | Random in Grid | | |
| -271.9052 | -111.7130 | | | Random in Grid | | |

**Primary Sampling Objective**
The primary purpose of sampling at this site is to find the proportion of sample locations on the site that have a certain characteristic of concern (such as presence of a contaminant above a specified level or detectable presence) and compute the upper confidence limit (UCL) on that proportion of samples. You can be 90% confident that the true proportion of samples on the site that have the characteristic of concern is below the UCL.

**Selected Sampling Approach**
A standard normal approximation of the binomial distribution was used to determine the number of samples because of the fluctuation in statistical power associated with the exact binomial proportion confidence interval. Small changes in the estimate of the proportion cause drastic changes in the associated statistical power (Chernick & Liu, 2002). This in turn translates to seemingly contradictory sample size requirements. Using the standard normal approximation guarantees that with greater uncertainty, the number of samples required to satisfy the confidence level increases. Given that the number of samples is modestly large and the proportion of unacceptable items is neither approximately 0 nor 1, then the sample size generated by the standard normal approximation is sufficient.

VSP offers many options to determine the locations at which measurements are made or samples are collected and subsequently measured. For this design, random point sampling in grids was chosen. This option offers a good balance between providing information about the spatial structure of the potential contamination while ensuring all portions of the site are represented (though, not as thoroughly as systematic grid sampling). Knowledge of the spatial structure is useful for geostatistical analysis. This option also has the benefit of placing the exact number of samples required by the design.

**Number of Total Samples: Calculation Equation and Inputs**
The equation used to calculate the number of samples is based on a standard normal approximation of the binomial distribution.

The formula used to calculate the number of samples is:

$$n = \frac{(z_{1-\alpha})^2 p(1-p)}{d^2}$$

where
$n$ is the number of samples required,
$\alpha$ is the maximum acceptable probability that the true proportion exceeds the UCL,
$z_{1-\alpha}$ is the value of the standard normal distribution such that the proportion of the distribution less than $z_{1-\alpha}$ is $1-\alpha$,
$d$ is the maximum desired difference between the estimated proportion and the UCL, and
$p$ is the maximum expected proportion. This proportion is set at 0.5 unless specified by the user.

The values of these inputs that result in the calculated number of sampling locations are:

| Analyte | n | Parameter | | | |
|---|---|---|---|---|---|
| | | d | $\alpha$ | $Z_{1-\alpha}$ [a] | p |
| Analyte 1 | 32 | 0.05 | 0.1 | 1.28155 | 0.05 |

[a] This value is automatically calculated by VSP based upon the user defined value of $\alpha$.

**Statistical Assumptions**

The assumptions associated with the formulas for computing the number of samples are:

1. the distribution of samples with the characteristic of concern follows a Binomial(n;p) distribution where n = total number of samples and p = proportion of unacceptable samples
2. the sampling locations will be selected randomly or any judgmentally selected samples are representative of the population. If using judgment sampling to select those locations where the likelihood of unacceptable samples is highest, the estimated proportion could be biased high. This may be acceptable if one desires an upper bound on the true proportion.

The these assumptions will be assessed in a post data collection analysis.

**Sensitivity Analysis**

The sensitivity of the calculation of number of samples was explored by varying the confidence level (1-$\alpha$) (%) and maximum difference between estimated proportion and ucl. The following table shows the results of this analysis.

| **Number of Samples** | | | |
|---|---|---|---|
| | **d=0.025** | **d=0.05** | **d=0.075** |
| **CL=94** | 184 | 46 | 21 |
| **CL=92** | 151 | 38 | 17 |
| **CL=90** | 125 | 32 | 14 |
| **CL=88** | 105 | 27 | 12 |
| **CL=86** | 89 | 23 | 10 |

CL = Confidence Level (1-$\alpha$) (%)
d = Maximum difference between estimated proportion and UCL

**Cost of Sampling**

The total cost of the completed sampling program depends on several cost inputs, some of which are fixed, and others that are based on the number of samples collected and measured. Based on the numbers of samples determined above, the estimated total cost of sampling and analysis at this site is $17,000.00, which averages out to a per sample cost of $531.25. The following table summarizes the inputs and resulting cost estimates.

| **COST INFORMATION** | | | |
|---|---|---|---|
| **Cost Details** | **Per Analysis** | **Per Sample** | **32 Samples** |
| Field collection costs | | $100.00 | $3,200.00 |
| Analytical costs (Analyte 1) | $400.00 | $400.00 | $12,800.00 |
| **Sum of Field & Analytical costs** | | **$500.00** | **$16,000.00** |
| Fixed planning and validation costs | | | $1,000.00 |
| **Total cost** | | | **$17,000.00** |

**Further Recommended Data Analysis Activities**

Post data collection activities generally follow those outlined in EPA's Guidance for Data Quality Assessment (EPA, 2000). The data analysts will become familiar with the context of the problem and goals for data collection and assessment. The data will be verified and validated before being subjected to statistical or other analyses. Graphical and analytical tools will be used to verify to the extent possible the assumptions of any statistical analyses that are performed as well as to achieve a general understanding of the data. The data will be assessed to determine whether they are adequate in both quality and quantity to support the primary objective of sampling.

**References**

Agresti, A., & Coull, B. (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions". *The American Statistician,* 52(2), 119-126.

Chernick, M. R., and Liu, C. Y. (2002). "The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods". *The American Statistician,* 56, 149-155.