Random sampling locations for comparing a mean with a fixed threshold (parametric)

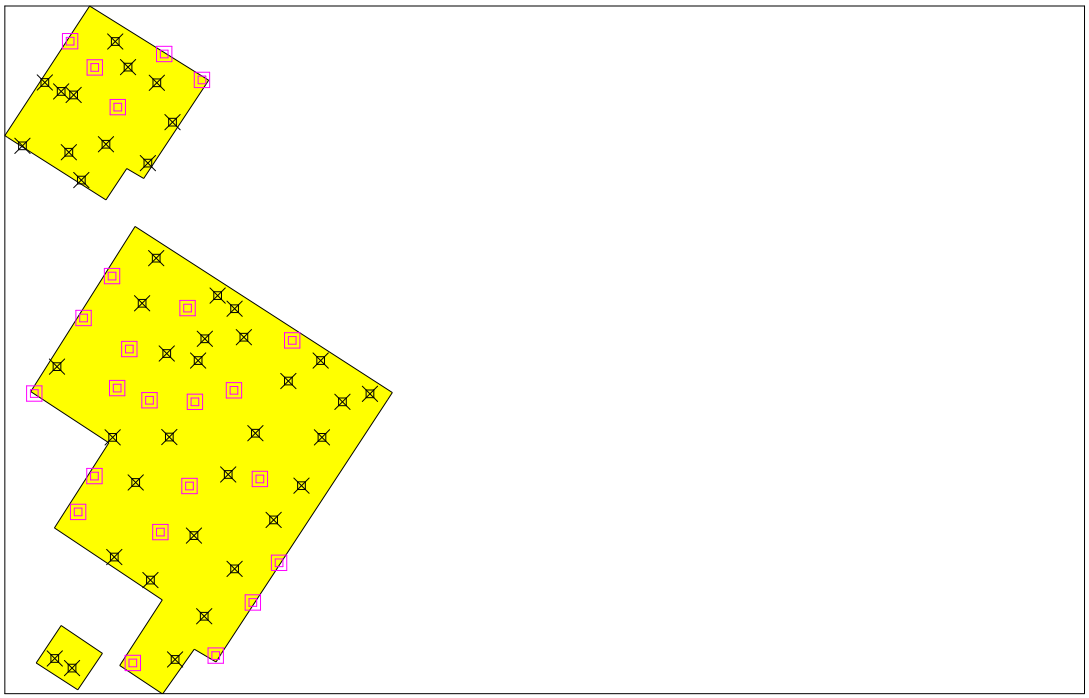
Summary
This report summarizes the sampling design, associated statistical assumptions, as well as general guidelines for conducting post-sampling data analysis. Sampling plan components presented here include how many sampling locations to choose and where within the sampling area to collect those samples. 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 the sampling plan.

The following table summarizes the sampling design. A figure that shows sampling locations in the field and a table that lists sampling location coordinates are also provided below.

SUMMARY OF SAMPLING DESIGN	
Primary Objective of Design	Compare a site mean to a fixed threshold
Type of Sampling Design	Parametric
Sample Placement (Location) in the Field	Simple random sampling
Working (Null) Hypothesis	The mean value at the site exceeds the threshold
Formula for calculating number of sampling locations	Student's t-test
Calculated total number of samples	65
Number of samples on map [a]	65
Number of selected sample areas [b]	2
Specified sampling area [c]	188054.34 m^2
Total cost of sampling [d]	$33,500.00

[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.

9557844



Area: Area 1					
X Coord	**Y Coord**	**Label**	**Value**	**Type**	**Historical**
679268.7700	3083200.3260	J-11S	0.0465	Manual	T
679301.1600	3083254.0340	J-12S	0.218	Manual	T
679171.2970	3083289.7960	J-04S	0.465	Manual	T
679155.0740	3083294.6960	J-03S	0.135	Manual	T
679133.4290	3083306.3130	J-01S	0.048	Manual	T
679104.2450	3083223.2620	J-02S	0.04	Manual	T
679164.8060	3083214.7100	J-06S	0.04225	Manual	T
679181.2750	3083178.2880	J-08S	0.0415	Manual	T
679213.7730	3083224.9730	J-09S	1.03	Manual	T
679280.5440	3083305.6810	J-10S	0.042	Manual	T
679242.7260	3083326.5280	J-07S	0.42	Manual	T
679225.8560	3083359.9740	J-05S	0.045	Manual	T
679339.9414	3083309.8490	J-15S	0.043	Adaptive-Fill	
679229.3268	3083274.1226	J-21S	0.44	Adaptive-Fill	
679166.5144	3083360.6002	J-19S	0.45	Adaptive-Fill	
679199.0125	3083325.9512	J-24S	0.41	Adaptive-Fill	
679290.1895	3083343.7836	J-20S	0.405	Adaptive-Fill	

Area: Area 3					
X Coord	**Y Coord**	**Label**	**Value**	**Type**	**Historical**
679261.0980	3083016.3510	J-15S	0.043	Manual	T
679335.0020	3082941.1720	J-21S	0.44	Manual	T

679343.5810	3082969.5980	J-19S	0.45	Manual	T
679394.8070	3082971.8300	J-24S	0.41	Manual	T
679382.8640	3083009.1130	J-20S	0.405	Manual	T
679293.5600	3082950.4980	J-17S	0.046	Manual	T
679360.5700	3083026.4980	J-18S	0.044	Manual	T
679297.0010	3082840.6970	J-23S	0.041	Manual	T
679252.7130	3082781.0290	J-22S	0.0425	Manual	T
679222.6340	3082840.1720	J-16S	0.0435	Manual	T
679279.6830	3083075.4290	J-14S	0.181	Manual	T
679149.4920	3082933.0980	J-13S	0.042	Manual	T
679272.0040	3082652.6750	J-28S	0.045	Manual	T
679224.5850	3082683.1400	J-26S	0.04275	Manual	T
679146.6460	3082549.7640	J-25S	0.04525	Manual	T
679169.0760	3082537.3510	J-27S	0.0445	Manual	T
679342.7410	3082605.3190	J-35S	0.0435	Manual	T
679304.6530	3082548.6880	J-34S	0.065	Manual	T
679382.8900	3082667.5270	J-36S	0.0455	Manual	T
679329.4380	3082711.0960	J-29S	0.0425	Manual	T
679374.4420	3082791.3300	J-30S	0.0485	Manual	T
679453.4760	3082914.1150	J-32S	0.048	Manual	T
679410.1490	3082845.8460	J-31S	0.04375	Manual	T
679560.6070	3082897.2580	J-41S	0.0385	Manual	T
679495.8840	3082940.9730	J-33S	0.042	Manual	T
679524.3310	3082886.8990	J-40S	0.043	Manual	T
679497.3310	3082840.3960	J-39S	0.0465	Manual	T
679470.3570	3082776.7350	J-38S	0.041	Manual	T
679433.9450	3082731.6820	J-37S	0.041	Manual	T
679184.4446	3082996.9264		0	Adaptive-Fill	
679177.1426	3082742.3877		0	Adaptive-Fill	
679228.5215	3082904.9556		0	Adaptive-Fill	
679381.8163	3082901.9806		0	Adaptive-Fill	
679441.0893	3082675.5360		0	Adaptive-Fill	
679358.0925	3082553.5465		0	Adaptive-Fill	
679330.4159	3082886.8562		0	Adaptive-Fill	
679221.7466	3083051.9971		0	Adaptive-Fill	
679198.5354	3082789.3512		0	Adaptive-Fill	
679249.1891	3082543.7262		0	Adaptive-Fill	
679244.2914	3082956.2636		0	Adaptive-Fill	
679406.6415	3082623.4861		0	Adaptive-Fill	
679119.5052	3082897.8954		0	Adaptive-Fill	
679271.0195	3082888.8959		0	Adaptive-Fill	

679458.5374	3082967.5965		0	Adaptive-Fill	
679285.1118	3082715.7475		0	Adaptive-Fill	
679320.7859	3083010.0230		0	Adaptive-Fill	
679323.6573	3082776.1377		0	Adaptive-Fill	
679415.7477	3082785.3452		0	Adaptive-Fill	

Primary Sampling Objective
The primary purpose of sampling at this site is to compare a mean value of a site with a fixed threshold. The working hypothesis (or 'null' hypothesis) is that the mean value at the site is equal to or exceeds the threshold. The alternative hypothesis is that the mean value is less than the threshold. VSP calculates the number of samples required to reject the null hypothesis in favor of the alternative hypothesis, given a selected sampling approach and inputs to the associated equation.

Selected Sampling Approach
A parametric random sampling approach was used to determine the number of samples and to specify sampling locations. A parametric formula was chosen because the conceptual model and historical information (e.g., historical data from this site or a very similar site) indicate that parametric assumptions are reasonable. These assumptions will be examined in post-sampling data analysis.

Both parametric and non-parametric approaches rely on assumptions about the population. However, non-parametric approaches typically require fewer assumptions and allow for more uncertainty about the statistical distribution of values at the site. The trade-off is that if the parametric assumptions are valid, the required number of samples is usually less than the number of samples required by non-parametric approaches.

Locating the sample points randomly provides data that are separated by many distances, whereas systematic samples are all equidistant apart. Therefore, random sampling provides more information about the spatial structure of the potential contamination than systematic sampling does. As with systematic sampling, random sampling also provides information regarding the mean value, but there is the possibility that areas of the site will not be represented with the same frequency as if uniform grid sampling were performed.

Number of Total Samples: Calculation Equation and Inputs
The equation used to calculate the number of samples is based on a Student's t-test. For this site, the null hypothesis is rejected in favor of the alternative hypothesis if the sample mean is sufficiently smaller than the threshold. The number of samples to collect is calculated so that 1) there will be a high probability (1-β) of rejecting the null hypothesis if the alternative hypothesis is true and 2) a low probability (α) of rejecting the null hypothesis if the null hypothesis is true.

The formula used to calculate the number of samples is:

$$n = \frac{S^2}{\Delta^2}\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 + 0.5Z_{1-\alpha}^2$$

where
n is the number of samples,
S is the estimated standard deviation of the measured values including analytical error,
Δ is the width of the gray region,
α is the acceptable probability of incorrectly concluding the site mean is less than the threshold,
β is the acceptable probability of incorrectly concluding the site mean exceeds the threshold,
$Z_{1-\alpha}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\alpha}$ is 1-α,
$Z_{1-\beta}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\beta}$ is 1-β.

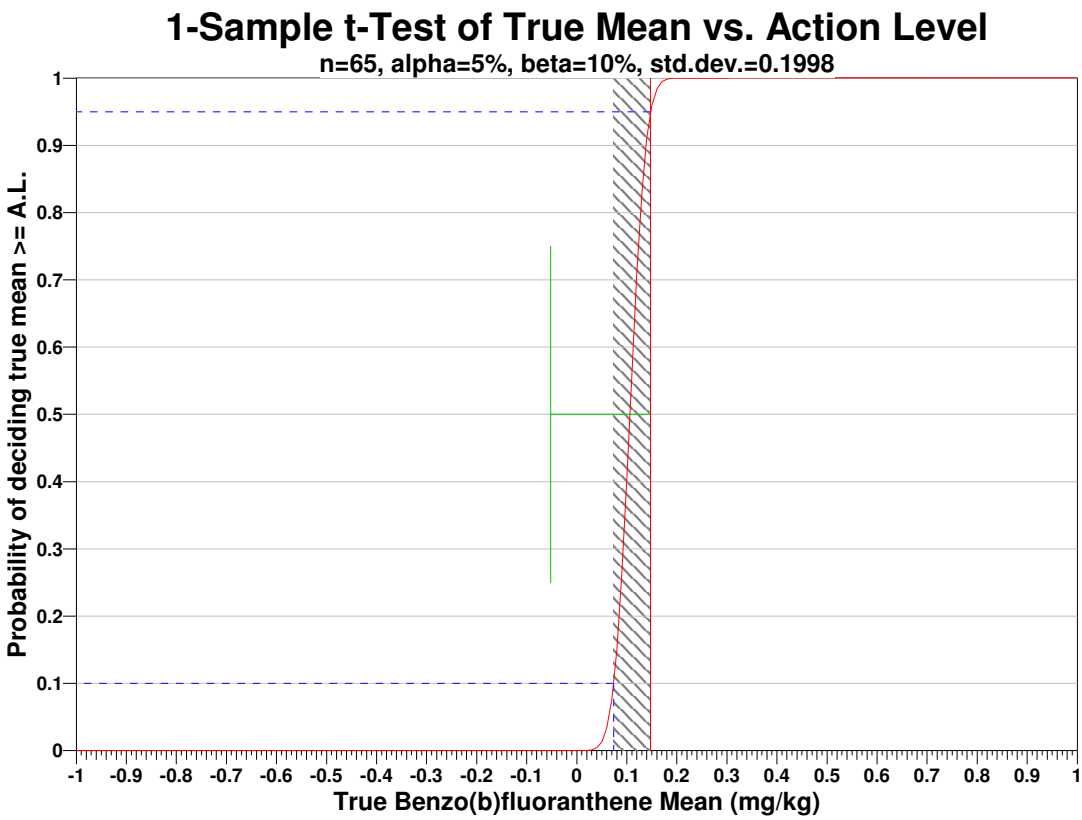
The values of these inputs that result in the calculated number of sampling locations are:

Analyte	**n**	**Parameter**					
		S	**Δ**	**α**	**β**	$Z_{1-\alpha}$ [a]	$Z_{1-\beta}$ [b]
Benzo(b)fluoranthene	65	0.1998 mg/kg	0.0738 mg/kg	0.05	0.1	1.64485	1.28155

[a] This value is automatically calculated by VSP based upon the user defined value of α.
[b] This value is automatically calculated by VSP based upon the user defined value of β.

The following figure is a performance goal diagram, described in EPA's QA/G-4 guidance (EPA, 2000). It shows the probability of concluding the sample area is dirty on the vertical axis versus a range of possible true mean values for the site on the horizontal axis. This graph contains all of the inputs to the number of samples equation and pictorially represents the calculation.

The red vertical line is shown at the threshold (action limit) on the horizontal axis. The width of the gray shaded area is equal to Δ; the upper horizontal dashed blue line is positioned at 1-α on the vertical axis; the lower horizontal dashed blue line is positioned at β on the vertical axis. The vertical green line is positioned at one standard deviation below the threshold. The shape of the red curve corresponds to the estimates of variability. The calculated number of samples results in the curve that passes through the lower bound of Δ at β and the upper bound of Δ at 1-α. If any of the inputs change, the number of samples that result in the correct curve changes.



Statistical Assumptions
The assumptions associated with the formulas for computing the number of samples are:
1. the sample mean is normally distributed (this happens if the data are roughly symmetric or the sample size is more than 30; for extremely skewed data sets, additional samples may be required for the sample mean to be normally distributed),
2. the variance estimate, S^2, is reasonable and representative of the population being sampled,
3. the population values are not spatially or temporally correlated, and
4. the sampling locations will be selected randomly.

The first three assumptions will be assessed in a post data collection analysis. The last assumption is valid because the sample locations were selected using a random process.

Sensitivity Analysis
The sensitivity of the calculation of number of samples was explored by varying the standard deviation, lower bound of gray region (% of action level), beta (%), probability of mistakenly concluding that μ > action level and alpha (%), probability of mistakenly concluding that μ < action level. The following table shows the results of this analysis.

Number of Samples

AL=0.1476		α=5		α=10		α=15	
		s=0.3996	s=0.1998	s=0.3996	s=0.1998	s=0.3996	s=0.1998
LBGR=90	β=5	7934	1985	6278	1571	5270	1318
	β=10	6279	1571	4816	1205	3939	986
	β=15	5271	1319	3940	986	3150	788
LBGR=80	β=5	1985	498	1571	394	1318	330
	β=10	1571	394	1205	302	986	247
	β=15	1319	331	986	247	788	198
LBGR=70	β=5	883	222	699	176	587	147
	β=10	699	176	536	135	439	110
	β=15	587	148	439	111	351	89

s = Standard Deviation
LBGR = Lower Bound of Gray Region (% of Action Level)
β = Beta (%), Probability of mistakenly concluding that μ > action level
α = Alpha (%), Probability of mistakenly concluding that μ < action level
AL = Action Level (Threshold)

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 $33,500.00, which averages out to a per sample cost of $515.38. The following table summarizes the inputs and resulting cost estimates.

COST INFORMATION			
Cost Details	**Per Analysis**	**Per Sample**	**65 Samples**
Field collection costs		$100.00	$6,500.00
Analytical costs	$400.00	$400.00	$26,000.00
Sum of Field & Analytical costs		**$500.00**	**$32,500.00**
Fixed planning and validation costs			$1,000.00
Total cost			**$33,500.00**

Data Analysis for Benzo(b)fluoranthene

SUMMARY STATISTICS for Benzo(b)fluoranthene	
n	2094
Min	0
Max	1.03
Range	1.03
Mean	0.0040505
Median	0
Variance	0.001489
StdDev	0.038588
Std Error	0.00084325
Skewness	15.668

Interquartile Range				0				
Percentiles								
1%	**5%**	**10%**	**25%**	**50%**	**75%**	**90%**	**95%**	**99%**
0	0	0	0	0	0	0	0	0.048

Outlier Test
Rosner's test for multiple outliers was performed to test whether the most extreme value is a statistical outlier. The test was conducted at the 5% significance level.

Data should not be excluded from analysis solely on the basis of the results of this or any other statistical test. If any values are flagged as possible outliers, further investigation is recommended to determine whether there is a plausible explanation that justifies removing or replacing them.

In using Rosner's test to detect up to 1 outlier, a test statistic R_1 is calculated, and compared with a critical value C_1 to test the hypothesis that there is one outlier in the data.

ROSNER'S OUTLIER TEST for Benzo(b)fluoranthene			
k	**Test Statistic R_k**	**5% Critical Value C_k**	**Significant?**
1	26.59	4.207	Yes

The test statistic 26.59 exceeded the corresponding critical value, therefore that test is significant and we conclude that the most extreme value is an outlier at the 5% significance level.

SUSPECTED OUTLIERS for Benzo(b)fluoranthene	
1	1.03

A normal distribution test indicated that the data do not appear to be normally distributed, so further investigation is recommended before using the results of this test. Because Rosner's test can be used only when the data without the suspected outlier are approximately normally distributed, a Lilliefors test for normality was performed at a 5% significance level.

NORMAL DISTRIBUTION TEST (excluding outliers)	
Lilliefors Test Statistic	0.5122
Lilliefors 5% Critical Value	0.01937

The calculated Lilliefors test statistic exceeds the 5% Lilliefors critical value, so the test rejects the hypothesis that the data are normal and concludes that the data, excluding the most extreme value, do not appear to follow a normal distribution at the 5% level of significance. Rosner's test may not be appropriate if the assumption of normally distributed data is not justified for this data set. Examine the Q-Q plot displayed below to further assess the normality of the data.

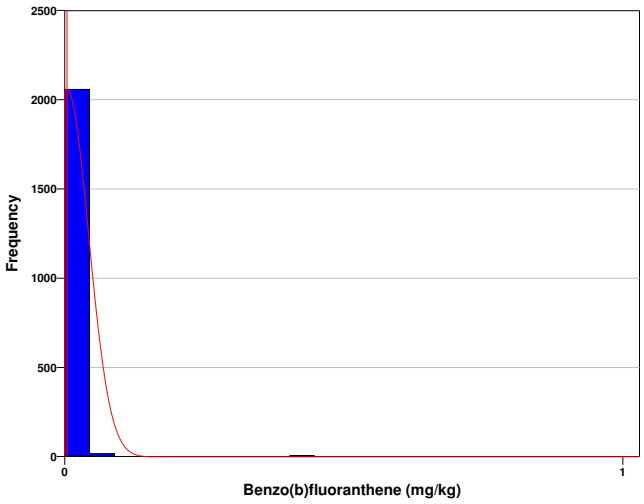
Data Plots for Benzo(b)fluoranthene
Graphical displays of the data are shown below.

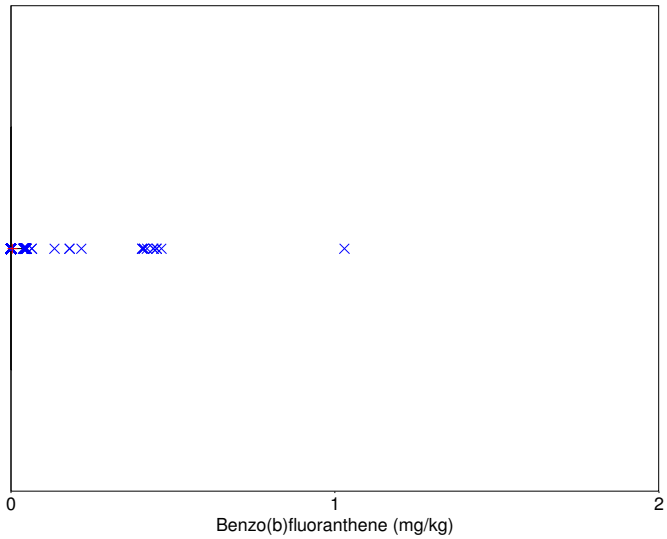
The Histogram is a plot of the fraction of the n observed data that fall within specified data “bins.” A histogram is generated by dividing the x axis (range of the observed data values) into "bins" and displaying the number of data in each bin as the height of a bar for the bin. The area of the bar is the fraction of the n data values that lie within the bin. The sum of the fractions for all bins equals one. A histogram is used to assess how the n data are distributed (spread) over their range of values. If the histogram is more or less symmetric and bell shaped, then the data may be normally distributed.

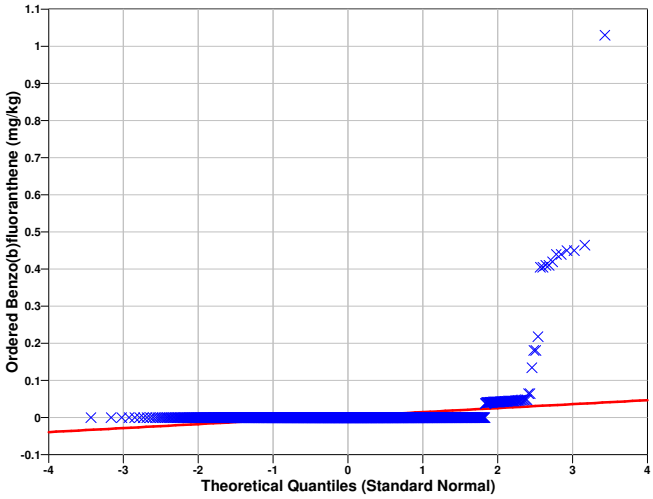
The Box and Whiskers plot is composed of a central box divided by a line, and with two lines extending out from the box, called the "whiskers". The line through the box is drawn at the median of the n data observed. The two ends of the box represent the 25th and 75th percentiles of the n data values, which are also called the lower and upper quartiles, respectively, of the data set. The sample mean (mean of the n data) is shown as a "+" sign. The upper whisker extends

to the largest data value that is less than the upper quartile plus 1.5 times the interquartile range (upper quartile minus the lower quartile). The lower whisker extends to the smallest data value that is greater than the lower quartile minus 1.5 times the interquartile range. Extreme data values (greater or smaller than the ends of the whiskers) are plotted individually as blue Xs. A Box and Whiskers plot is used to assess the symmetry of the distribution of the data set. If the distribution is symmetrical, the box is divided into two equal halves by the median, the whiskers will be the same length, and the number of extreme data points will be distributed equally on either end of the plot.

The Q-Q plot graphs the quantiles of a set of n data against the quantiles of a specific distribution. We show here only the Q-Q plot for an assumed normal distribution. The pth quantile of a distribution of data is the data value, x_p, for which a fraction p of the distribution is less than x_p. If the data plotted on the normal distribution Q-Q plot closely follow a straight line, even at the ends of the line, then the data may be assumed to be normally distributed. If the data points deviate substantially from a linear line, then the data are not normally distributed.







For more information on these plots consult Guidance for Data Quality Assessment, EPA QA/G-9, pgs 2.3-1 through 2.3-12. (http://www.epa.gov/quality/qa-docs.html).

Tests for Benzo(b)fluoranthene

A goodness-of-fit test was performed to test whether the data set had been drawn from an underlying normal distribution. The Lilliefors test was used to test the null hypothesis that the data are normally distributed. The test was conducted at the 5% significance level, i.e., the probability the test incorrectly rejects the null hypothesis was set at 0.05.

NORMAL DISTRIBUTION TEST	
Lilliefors Test Statistic	0.5084
Lilliefors 5% Critical Value	0.01936

The calculated Lilliefors test statistic exceeds the 5% Lilliefors critical value, so we can reject the hypothesis that the data are normal, or in other words the data do not appear to follow a normal distribution at the 5% level of significance. The Q-Q plot displayed above should be used to further assess the normality of the data.

Upper Confidence Limit on the True Mean
Two methods were used to compute the upper confidence limit (UCL) on the mean. The first is a parametric method that assumes a normal distribution. The second is the Chebyshev method, which requires no distributional assumption.

UCLs ON THE MEAN	
95% Parametric UCL	0.005438
95% Non-Parametric (Chebyshev) UCL	0.007726

Because the data do not appear to be normally distributed according to the goodness-of-fit test performed above, the non-parametric UCL (0.007726) may be a more accurate upper confidence limit on the true mean.

One-Sample t-Test
A one-sample t-test was performed to compare the sample mean to the action level. The null hypothesis used is that the true mean equals or exceeds the action level (AL). The t-test was conducted at the 5% significance level. The sample value *t* was computed using the following equation:

$$t = \frac{\bar{x} - AL}{SE}$$

where
x is the sample mean of the n=2094 data,
AL is the action level or threshold (0.1476),
SE is the standard error = (standard deviation) / (square root of n).

This *t* was then compared with the critical value $t_{0.95}$, where $t_{0.95}$ is the value of the t distribution with n-1=2093 degrees of freedom for which the proportion of the distribution to the left of $t_{0.95}$ is 0.95. The null hypothesis will be rejected if $t < -t_{0.95}$.

ONE-SAMPLE t-TEST		
t-statistic	**Critical Value $t_{0.95}$**	**Null Hypothesis**
-170.23	1.6456	Reject

The test rejected the null hypothesis that the mean value at the site exceeds the threshold, therefore conclude the true mean is less than the threshold.

Because the data do not appear to be normally distributed, the MARSSIM Sign Test might be preferred over the One Sample t-Test. The following table represents the results of the MARSSIM Sign Test using the current data:

MARSSIM Sign Test		
Test Statistic (S+)	**95% Critical Value**	**Null Hypothesis**
2080	1085	Reject

This report was automatically produced* by Visual Sample Plan (VSP) software version 5.3.1.
Software and documentation available at http://dqo.pnl.gov/vsp
Software copyright (c) 2009 Battelle Memorial Institute. All rights reserved.
* - The report contents may have been modified or reformatted by end-user of software.