

Chapter 12: Statistics

Once you have imported your data or created a geospatial model, you may wish to calculate some simple statistics, run some simple tests, or see some traditional plots. On the main menu, you'll find the Statistics menu item which contains most of the standard statistical features provided.

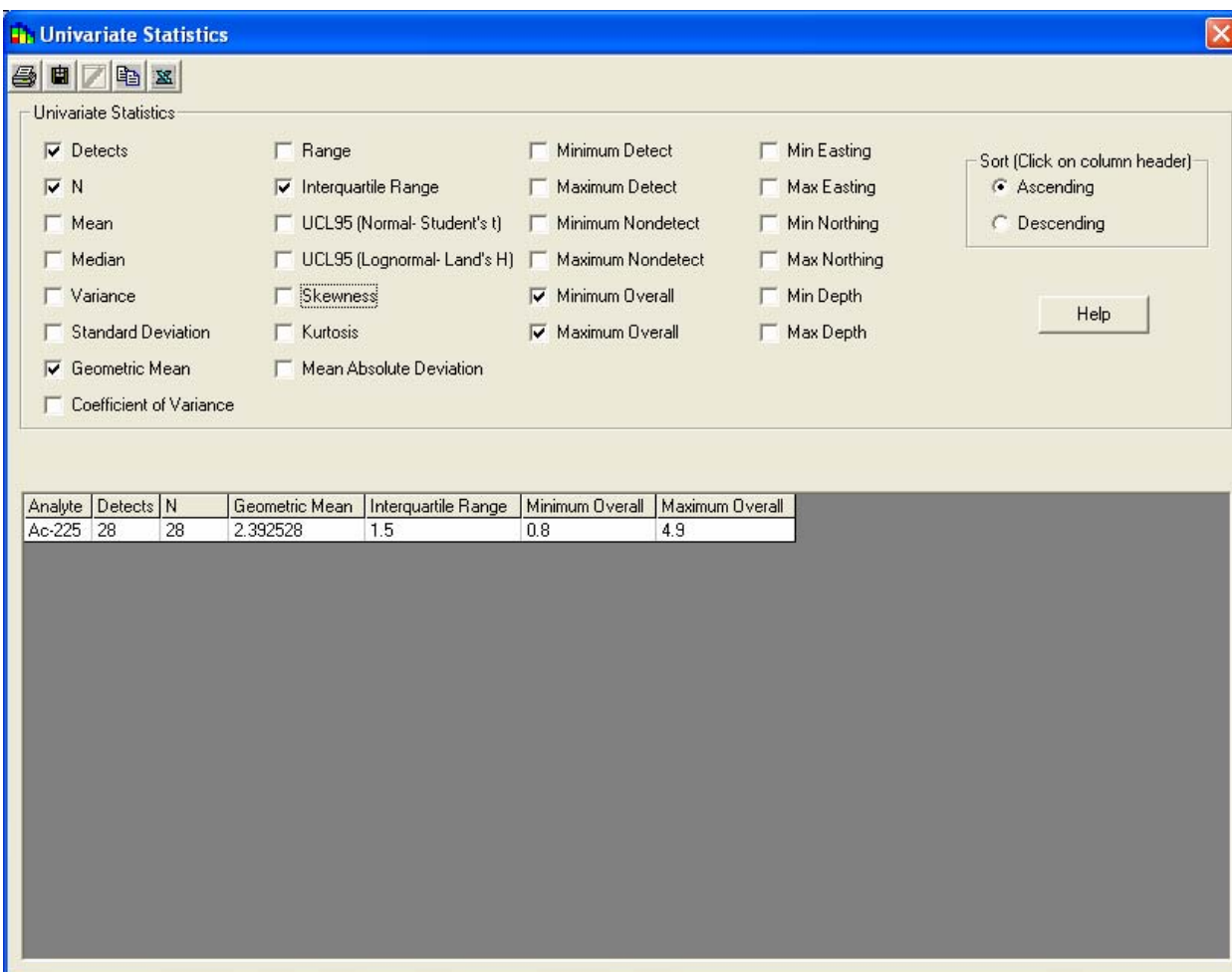


Let's begin by opening the file Statistics.sda (File→Open and navigate to where SADA is installed). SADA opens up your file with the Soil measurements for Ac-225.

Univariate

This feature will show you a collection of standard statistical endpoints such as mean, variance and so forth for whatever you currently have selected. In this case, we have Ac-225 selected. Choose Statistics→Univariate and the following window is presented. You can alternatively press the statistics button on the toolbar.





You can choose which statistical values you want to see by selecting or deselecting them in the top half of the window. Also you can sort columns by different values. Simple click on the column you want to sort and SADA will sort them according to your sort selection found on the right hand side. Continue reading for details on each statistical endpoint.

Detects

The number of total samples that were detected, presented in the form (detected/total N)

N

The total number of values.

Mean

The sum of the values of a variable divided by the number of values

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Median

The value below which 50% of the data values fall

odd n $\tilde{X} = X_{([n+1]/2)}$

even n $\tilde{X} = \frac{X_{(n/2)} + X_{([n/2]+1)}}{2}$

Variance

A parameter that measures how dispersed a random variable's probability distribution is, the mean of the squares of the differences between the respective samples and their mean

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Standard Deviation

The positive square root of the variance

$$S = \sqrt{S^2}$$

Geometric Mean

The mean of n numbers expressed as the n-th root of their product .

$$\exp \left[\frac{1}{n} \sum_{i=1}^n \ln x_i \right]$$

Coefficient of Variance

The ratio of the variance over the mean

$$CV = s / \bar{X} = \frac{[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2]^{1/2}}{\frac{1}{n} \sum_{i=1}^n X_i}$$

Range

The difference between the lowest and highest values

$$R = \text{maximum} - \text{minimum}$$

Interquartile Range

The central portion of a distribution, calculated as the difference between the third quartile and the first quartile; this range includes about one-half of the observations in the set

$$IQR = y(75) - y(25)$$

UCL95 (Normal – Student's t)

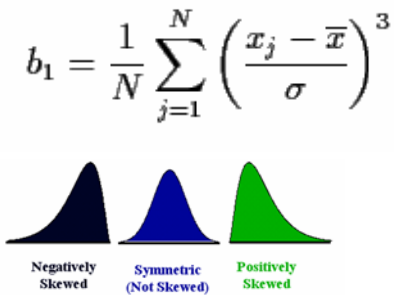
Upper 95% confidence limit on the mean concentration of a normal distribution

UCL95 (Lognormal – Land's H)

Upper 95% confidence limit on the mean concentration of a lognormal distribution

Skewness

A measure of symmetry, skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right



Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

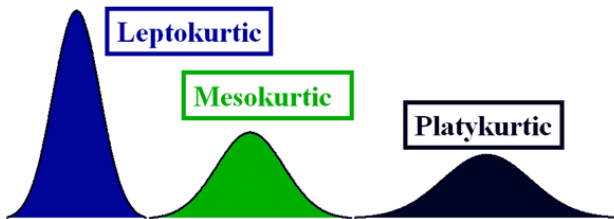
$$b_2 = \frac{1}{N} \sum_{j=1}^N \left(\frac{x_j - \bar{x}}{\sigma} \right)^4$$

Kurtosis can be grouped into three broad categories

Leptokurtic: high and thin (high kurtosis values)

Mesokurtic: normal in shape

Platykurtic: flat and spread out (low kurtosis values)



Mean Absolute Deviation

The mean of the absolute values of the differences between the respective samples and their mean

$$M. A. D. = \frac{\sum |X - \mu|}{N}$$

Minimum Detect

The lowest detect value found in the dataset.

Maximum Detect

The highest detected value found in the dataset.

Minimum Non-detect

The lowest detect value found in the dataset

Maximum Non-detect

The highest detected value found in the dataset

Minimum Overall

The lowest detect value found in the dataset

Maximum Overall

The highest detected value found in the dataset

Min Easting

Smallest easting coordinate value

Max Easting

Largest easting coordinate value

Min Northing

Smallest northing coordinate value

Max Northing

Largest northing coordinate value

Min Depth

Smallest depth below surface value.

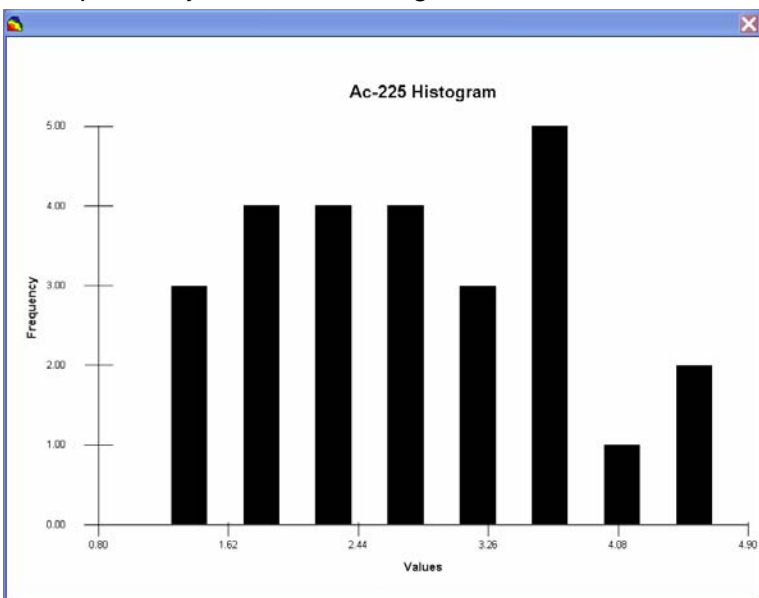
Max Depth

Largest depth below surface value

Obviously, certain univariate statistics would have little meaning when applied to a set of model values. For example, the upper confidence limit on the mean is meaningless since it is a function of the number of values and the number of cells in a grid can be easily varied.

Show Histogram

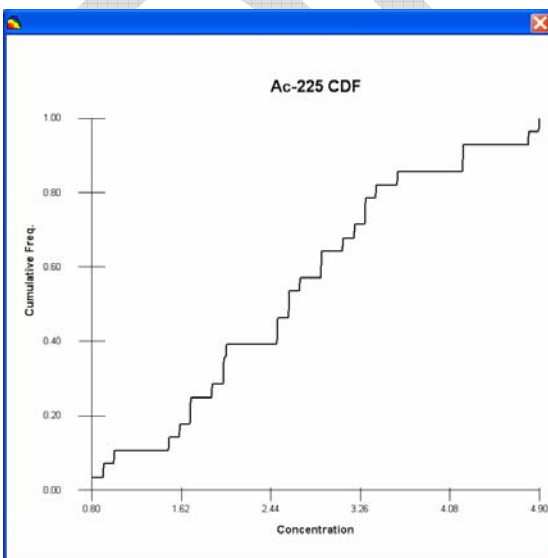
Looking at the histogram of values for either data or model results is very valuable in getting a visual sense for the distribution of your results. The histogram is always shown for the current result in your map. Of course histograms are not available for everything that may appear in your viewer (e.g. cost benefit analysis). The menu item Statistics → Show Histogram will show present you with the histogram view. Select this menu now.



The Show Histogram menu actually acts as a toggle. Select it again to switch back to your spatial view.

Show CDF

The cumulative distribution function or CDF is a natural companion to the histogram and is focus of evaluation for certain geospatial models (e.g. indicator kriging). To see the CDF select Statistics→Show CDF and the following result appears in your viewer for Ac-225.

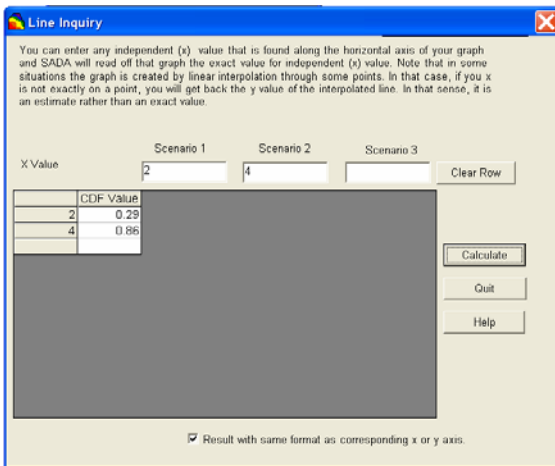


On the x-axis you see the range of Ac-225 concentration values. On the y-axis is the percentage or more accurately the proportion of data values that fall below a given concentration value. So for an x-axis reading of 0.80 we have no data that fall below this value since this is the smallest Ac-225 reading in the data set. The top reading of 4.90 pCi/g shows a value near 1.0 as this is the highest value in the data set.

You can use the Line Query tool button to directly read up to 3 values from the line. Select this button now.



You will be presented with this window.



Here Scenario just means a particular x value. Enter 2 and 4 into Scenario 1 and Scenario 2 respectively. Press Calculate and SADA calculates that the proportion of sample values below 2 is 0.29 and 4 is 0.86. Note that you can match the format of reported CDF values with the y axis in the CDF plot by checking the option at the bottom of the window. The line query tool can be used with cost benefit plots as well.

Press the Quit button.

Statistical Tests

A statistical test is a procedure for deciding whether a hypothesis about a population of values is true or false. For example, one might hypothesize that the mean of the population is less than 5 pCi/g. Of course, the only way to know for certain is to have access to the entire population of values. Here the population would be the exhaustive set of Ac-225 measurements. That is, a measured value of Ac-225 for every single location on the site. This of course is impossible. Instead only a sample of the population is available. Therefore statistical tests are used to draw a reasonable conclusion about the hypothesis when the entire data set is not available.

Furthermore, statistical tests are instrumental in separating significant effects from mere luck or random chance. For example, suppose that you flipped a coin twice landing on tails both times. We know that a fair coin should see heads 50% of the time. With so few flips (only 2) we cannot conclude whether the coin is fair or not. We might have tails two times in a row by mere luck. Or it might be true that the coin is biased so badly we would get tails every single time. A statistical test would inform us that we cannot separate out luck from true bias in this situation. However, after 100 flips a test could discern the difference. This is the power of a statistical test.

Of course, without the full population at hand, any conclusion drawn is at risk for error. Fortunately this risk is quantifiable. Two main types of error can occur:

1. A type I error occurs when a true hypothesis is rejected (a false negative in terms of the null hypothesis).
2. A type II error occurs when a false hypothesis is accepted (a false positive in terms of the null hypothesis).

Decision Based on Sample Data	True Condition	
	Baseline is True	Alternative is True
Decide baseline is true	Correct Decision	<i>Decision Error (False Acceptance)</i>
Decide alternative is true	<i>Decision Error (False Rejection)</i>	Correct Decision

*Taken from Guidance for the Data Quality Objectives Process USEPA 2000 QA/G-5, EPA/600/R-96/055

Officially speaking, a null hypothesis is a statistical hypothesis that is tested for possible rejection under the assumption that it is true (usually that observations are the result of chance). The alternative hypothesis is the hypothesis contrary to the null hypothesis.

The process for a statistical test is broadly conducted as follows. First, form a hypothesis about some statistical endpoint (e.g. the mean < 5) of your population (which will likely never have). This is your null hypothesis and the test will attempt to reject it by evaluation of the data. Next draw a random sample from the population (this is the Ac-225 data you have currently selected). Typically an assumption about the underlying distribution is made (e.g. normally distribution). Using this assumption choose an appropriate statistical test and compare the result to a critical level.

There are two broad types of statistical tests: parametric and non-parametric. A parametric test makes an assumption about the underlying distribution of observed data. A non-parametric test makes no such assumption.

In particular, non-parametric tests do not assume normal distributions, can handle non-detects (which often cause truncated distribution shapes), are insensitive to outliers and work nearly as well as their parametric counterparts when applied to normally distributed data.

SADA currently implements two non-parametric tests used by the DQO and MARSSIM processes: the Sign Test and the Wilcoxon Rank Sum Test.

Whether parametric or nonparametric, there are two different types of test you can conduct: two sided and one sided. A two sided hypothesis states that there is a difference between the two groups (or values) being tested, but does not specify in advance what direction this difference will be. In an environmental context, one group might be Ac-225 measurements in the contaminated area. The second group could be Ac-225 measurements in background. They hypothesis could be that there is no difference between the two groups. A one sided hypothesis states a specific direction (e.g., the site concentrations are greater than the reference site concentrations).

We now present the two nonparametric tests that SADA provides: Sign and Wilcoxon Rank Sum (WRS) Test.

Sign Test

In a sign test, we do the following. First take take the difference between each measured value and the decision criteria. Some measured values will be less than the criteria and will produce a negative difference. Some measured values will be greater than the criteria and will produce a positive difference. Some measured values may be exactly equal to the decision criteria producing a difference of zero. In the following example, we've taken some Arsenic measurements and calcluated the difference between them and a decision criteria of 10mg/kg.

Arsenic	Criteria	Difference	Sign
12	10	2	+
28	10	18	+
8	10	-2	-
42	10	32	+
16	10	6	+
23	10	13	+
45	10	35	+
31	10	21	+

The sign test, true to its name, only cares about whether the sign of the result is positive or negative as seen in the last column. How large or small the difference is becomes irrelevant. What the sign test does is test whether the number of +'s and -'s are equal. Because of this formulation, the sign test uses wording in the null and alternative hypotheses that is a bit unintuitive so let's take some time with it. The null and alternative hypothesis (Gilbert, 1987, p.242) are written as follows.

H0: the median of the population of all possible differences is zero.

H1: the median of the population of all possible difference does not equal zero.

Let's stop for a moment and understand what is being said here in less formal language. Let's start with the "population of all possible differences." If you could take ever single Arsenic measurement of the site you would have the population of Arsenic values. If you then subtracted them from the decision criteria (10mg/kg) then note the number of +'s and -'s, you would move yourself from the population of measurements to the population of differences.

In the table above, we are subtracting measured values from the decision criteria and ending up with differences from which we record the number of +’s and –’s. So we move ourselves from the sample of “measured values” to a sample of “differences”.

Why do we care about the median of these differences? If the population (not the sample) of differences if there exactly the same number of +’s as there are –’s then we have central or median value of zero. Think of these as +1s and -1s and if you have the same number then the median of these would be halfway between or zero. Note that if you end up with any difference exactly zero, it is thrown out (Gilbert, 1987).

So what does this have to do with determining if our site is contaminated or not? Let’s start by reformulating the hypothesis into an equivalent statement.

H0: Measured values are just as likely to exceed the criteria as they are to be less than the criteria.

H1: One of the following is true:

- A) measured values are more likely to exceed the criteria
- B) measured values are more likely to be less than the criteria

These are equivalent to the first formulation. In environmental assessment, we are more likely to be interested in H1A. We could then rewrite this as the following.

H0: Measured values are just as likely to exceed the criteria as they are to be less than the criteria.

H1: Measured values are more likely to exceed the criteria

For the example we show above let’s write this one more time very specifically.

H0: Arsenic values are just as likely to exceed 10mg/kg as they are to be less than 10mg/kg.

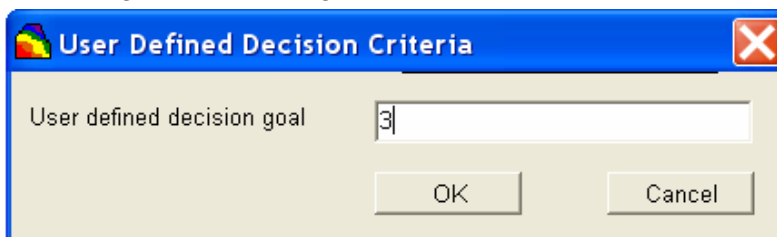
H1: Arsenic values are more likely to exceed 10mg/kg

The critical value used to conduct this one sided test is provided as the number of +s that you must have to reject the null and it comes from the binomial distribution. The sign test critical value tables can be found in introductory statistics books. In SADA it is computed directly. Let’s test the hypothesis now.

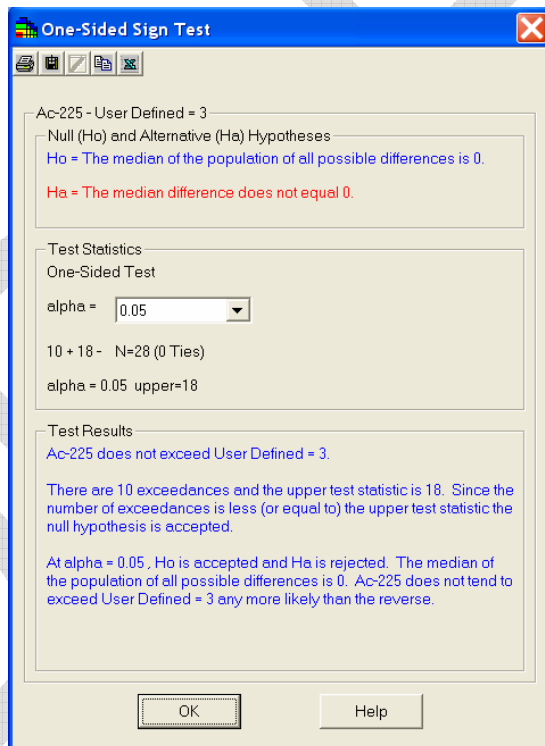
In our example, we want to be very sure that if do reject the null hypothesis, we want there to be only a 5% chance we did the wrong thing. This is the type I error we talked about earlier. This value is referred to as alpha (so we have $\alpha = .05$). The next thing we need is the test statistic B and a critical value to compare it to for $\alpha = .05$.

In our example, counting up the number of samples yields $N=8$ and the number of exceedances is $B=7$. Now turning to the binomial distribution table (most statistics books have the table in the back) for a one-sided sign test with $\alpha = .05$, $N=8$, and $B = 7$ we see the critical value is also 7. Therefore, since our test statistic equals the critical value we are led to reject the null hypothesis. Since we computed the number of exceedances as our test statistic, this would in fact indicate the median is exceeding 10 mg/kg. This means that more often than not, Arsenic values are exceed the decision criteria. When we apply our decision rule, we infer that the site is contaminated.

To use the sign test in our current file, select from the main menu Statistics→Statistical Tests→ Sign Test vs Decision Criteria. If you are working under the General Analysis you will be presented with the following window asking for a decision criteria.



The human health, ecological, and custom analyses would present their usual windows for choosing decision criteria. These will be covered a little later. Enter a value of 3 here (pCi/g) and press Ok. The sign test window is presented. At the top we have the null and alternative hypothesis. The two-sided alternative is written there but the one sided formulation would have been more appropriate. Look for updates as SADA releases continue. Nevertheless, we are doing a one sided upper test where the HA is that the median exceeds zero.



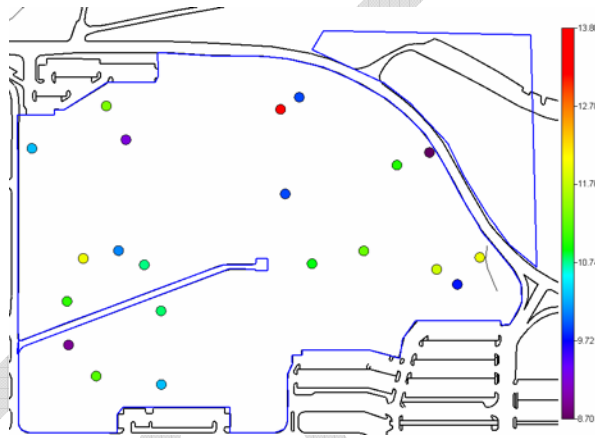
So at alpha = 0.05, with N=28, a decision criteria of 3pCi/g, and a test statistic of 18, we cannot conclude that the median of Ac-225 values exceeds 3 because we only have 10 exceedances.

Wilcoxon Rank Sum Test (WRS)

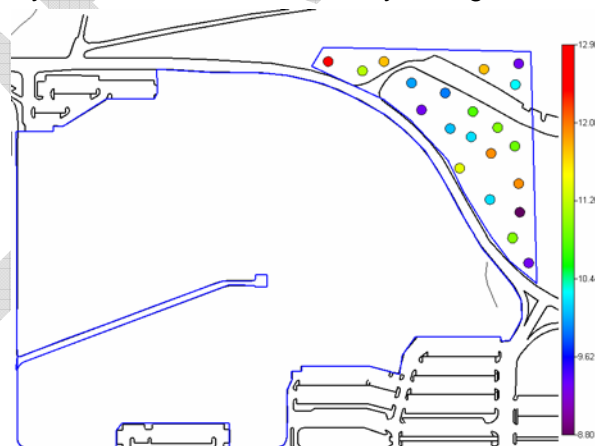
The WRS is another non-parametric test that is implemented as either a one-sided or a two-sided test in SADA. In this case you compare one contaminant data set versus another data set (e.g. background).

Specifically, the WRS test uses a sum of ranks comparison to determine if the two data sets have different means. The process is conducted as follows. First combine data sets and order from lowest to highest, each data value receives a rank (e.g. 1st, 2nd, 3rd, etc). Sum the ranks for the two different populations and compute the WRS test statistic (different forms depending on if there are ties) Compare to critical value for m and n sample sizes. Null hypothesis is either accepted or rejected.

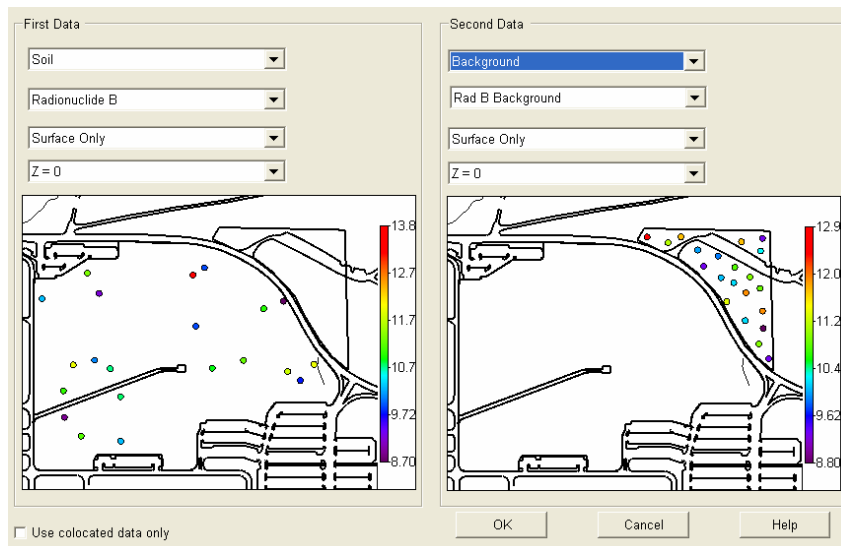
To see how this works, first open the file RadWithBackground.sda (you can save your current file or not). This SADA file contains some synthetic radiological data along with some background measurements. With Soil selected, choose Radionuclide B. You should see the following map.



To see the background data, select Background instead of Soil. The Background data for Radionuclide B will display since it is the first and only background dataset in this file.



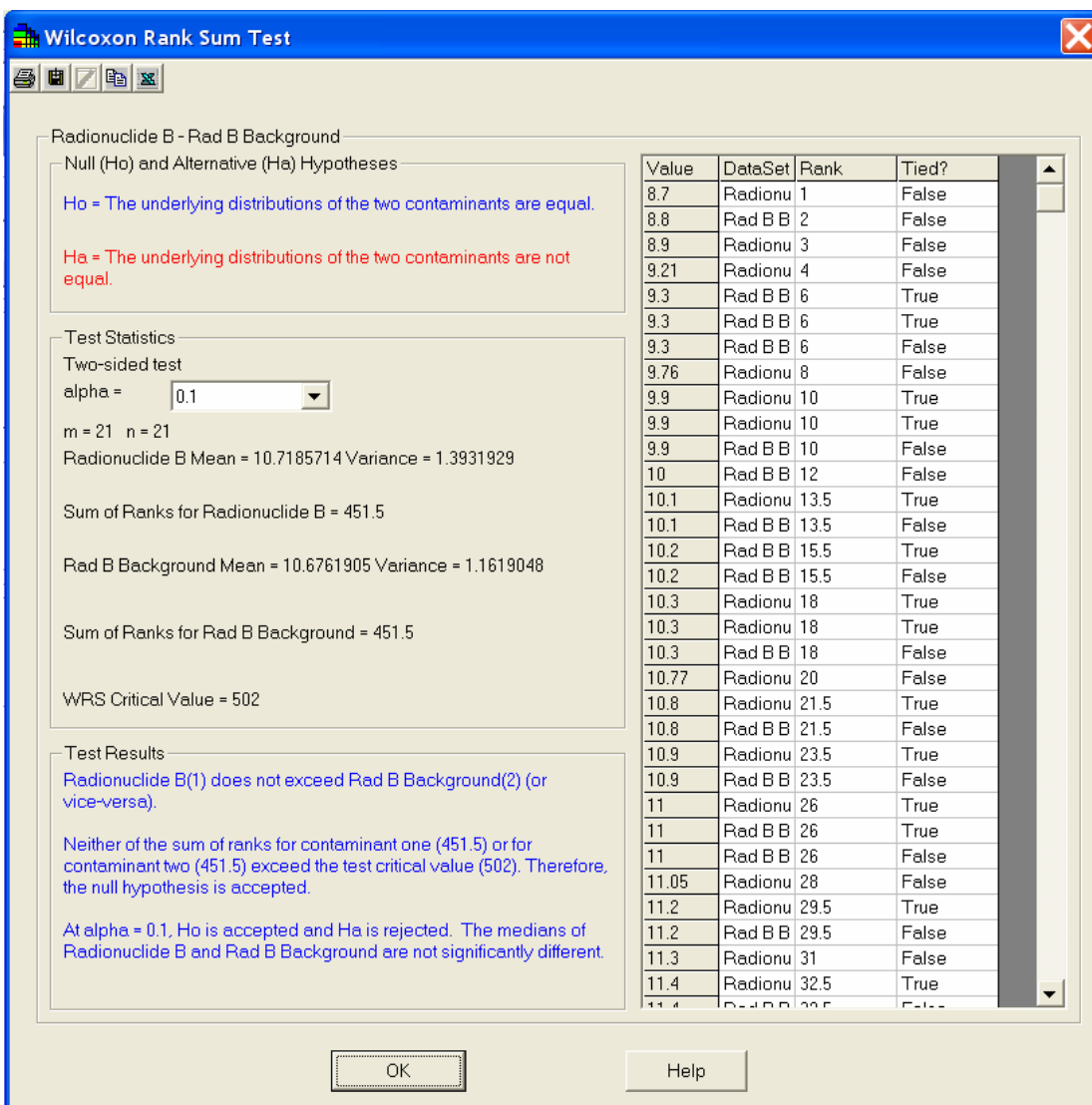
Next we will perform a WRS test to see if the site data is greater than the background. To do this, select from the main menu Statistics→Statistical Tests→Wilcoxon Rank Sum and the WRS test feature is initiated with following window. Select in this window the two data sets you want to compare. In this case, select Radionuclide B from the left window and Radionuclide B Background from the right window. SADA shows a preview of each.



Two important points should be brought to your attention here. First in the lower right hand corner we have the option to only use collocated data. There are some situations where data occur concurrently and this may be appropriate. In this example, where we are comparing site and background data we have no collocated data and so this is not appropriate.

Secondly, notice that you can select layering schemes as well. This is important. You may want to setup a particular layering scheme with polygons embedded in such a way as to include or exclude a set of data. For example, you may have background data from two locations and wish to exclude one while you use the other. In the other dataset selection, it a different layering/polygon scheme may be appropriate. This permits you the flexibility to spatially zero in on two or more different locations when doing the WRS test. In this example, we will use the Surface Only layer with no polygons other than our site boundary polygons.

Press Ok.



Here the null hypothesis is that the two groups are the same (site is no different than background). The alternative is that they are different. At the alpha value of .1 We accept the null hypothesis that the two populations are different and might conclude the site is clean. Had the null been rejected, SADA would have also conducted a one sided test to determine if the site data in fact exceeds background.

Number of Samples

This topic is handled in the chapter on Overview of Sample Designs

MARSSIM Quick Calculations

This topic is handled in the MARSSIM chapter.

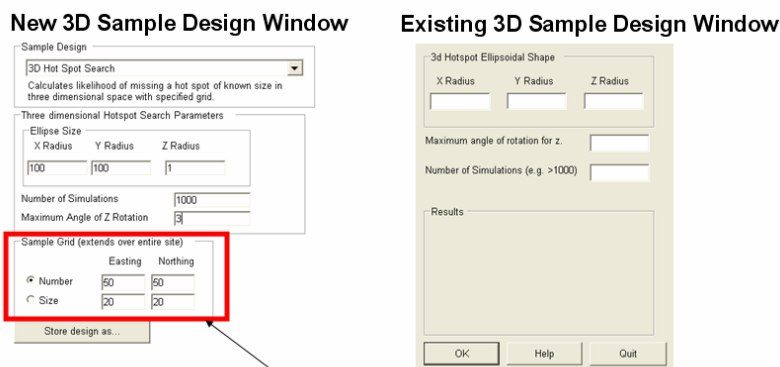
3d Search Efficiency

Among SADA's many sampling strategies is the 3d Hot Spot search design where a 3d grid of candidate locations is positioned on the site and the probability of discovering a particular 3d volume is calculated. In this feature, we generalized this to any sample design, gridded or not and apply it to *existing* sample designs. In other words, to use this feature, you will have needed to import your data first. There are a couple of situations in which this feature can be useful.

First, a 3d search grid was implemented in the past, but when the samples were taken adjustments had to be made due to physical obstructions or cost. This feature can analyze your resulting design and see how the probability of discovery might be affected.

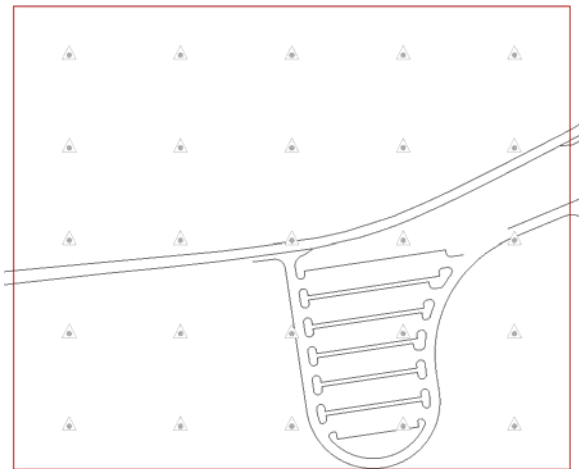
Secondly, the data you have may originate from a sample design that was not primarily interested in searching for a volume of a givens size. However, now the issue has arisen and the interest is in evaluating how well it serves the 3d discovery objective.

The details of the feature are found in the chapter on initial sample designs under the 3d Hot Spot Search discussion. You are encouraged to read that material first. We will point out here that the only difference in the interface is under the sample design strategy you will enter a candidate 3d grid which is then evaluated. Under this feature, the currently selected design is evaluated.

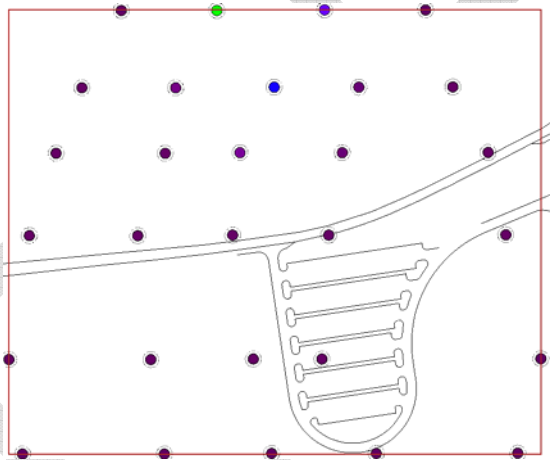


The initial sample design proposed and evaluated here is replaced by the actual sample locations in the 3d Search Efficiency Feature

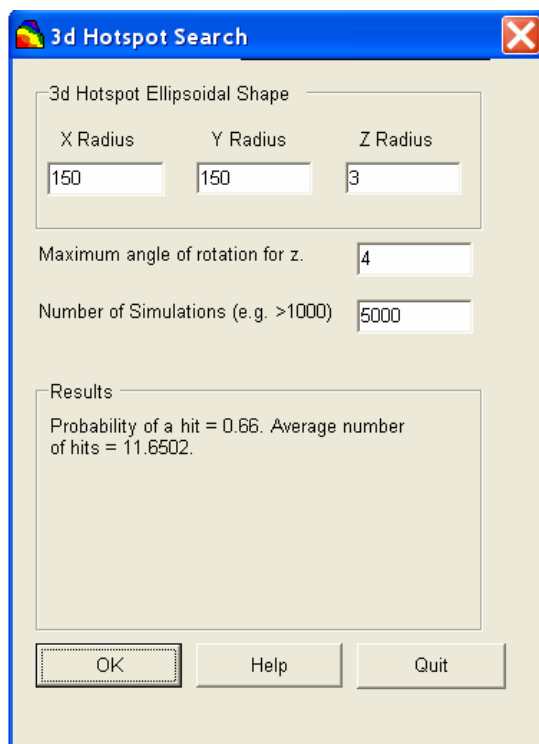
For a quick demonstration, open the file ThreeDimensional.sda. The is a 3d data set that was originally a 5 x 5 3d design with a 74% probability of discovering an elliptical 3d volume of dimensions 150 x 150 x 3ft (see initial design chapters for definition of 3d volume geometry).



However, when the samples were collected they could not adhere to this strict gridded pattern. As you can see by the Chlordane measurements that adjustments were made resulting in the following map.



In this hypothetical example, two things are observed. First the deviation from the strict sampling grid. Secondly suppose that a screening of the values yielded no points above the decision criteria (see chapters on screening criteria). The question becomes, could we have missed an elevated area with more likelihood than 26% ($100-74=26\%$)? To find out, select from the Statistics menu 3d Hotspot Search efficiency. Enter the parameters as you see them below and Press Ok.



Here we see in the Results summary that the probability of a hit (or probability of discovering an elevated ellipsoidal volume) is now about 66% (your value may vary slightly due to the nature of random realizations/simulations). This represents about a 8% ($74-66=8\%$) reduction in reliability due to the adjustments made at sample time. The average number of hits is around 11.6. This means that on average, when a 3d ellipsoid is simulated hit is “discovered” or “hit” by about 11 existing samples each time. Press the Quit button to close the window.

Summary

SADA provides a basic set of statistical features. If you are in need other statistics, you can export your data out of SADA into a csv file where it can then be imported into a statistics package. Finally, the many of these statistical features apply to both gridded and point data. It is important particularly in the case of some univariate statistics to properly interpret gridded statistics. For example, SADA will report a UCL95 for a modeled dataset. This value is meaningless however as it depends on the number of values which is arbitrarily set by the modeler. For an overview of environmental statistics we find the book by Richard Gilbert, Statistical Methods for Environmental Pollution Monitoring to be an excellent source.

References

Gilbert, R. 1987 Statistical Methods for Environmental Pollution Monitoring, Wiley & Sons, NY.