Analyzing Datasets Containing Semi-quantitative Values

Instructors

John Warren
Office of Environmental Information, EPA

Barry Nussbaum
Office of Environmental Information, EPA

2

Schedule

1:00	Introduction
1:15	Dealing with Censored Data
2:00	Using the Kaplan-Meier Method (class exercise)
2:30	Comparison of Methods and Conclusion
2:45	*Break*
3:00	Using Visual Sample Plan
4:30	*Conclude*

3

Purpose of the presentation

- To consider several different methods for dealing with censored data:
  - replace <DL with zero
  - replace <DL with DL
  - replace <DL with DL/2
  - adjust estimators using Normality (Cohen)
  - estimate directly using Normality (Regression)
  - estimate directly nonparametrically (Kaplan-Meier)
- To gain an understanding of how Visual Sample Plan can be used to analyze data already collected

4

Dealing with Censored Data

Purpose of the talk

- To consider several different methods for dealing with censored data:
  - replace <DL with zero
  - replace <DL with DL
  - replace <DL with DL/2
  - adjust estimators using Normality (Cohen)
  - estimate directly using Normality (Regression)
  - estimate directly nonparametrically (Kaplan-Meier)

Arsenic in Soil Samples

- **The following data values have been recorded and submitted for analysis and comment:**

1.752	<1.000 (0.944)	1.418	1.477	<1.000 (0.897)
<1.000 (0.777)	1.289	1.498	<1.000 (0.818)	<1.000 (0.622)
1.327	<1.000 (0.919)	1.060	<1.000 (0.455)	1.358
<1.000 (0.977)	1.045	<1.000 (0.806)	1.563	1.148

(Actually the method is accurate to 3 decimal places and can measure reliably down to its own limit of 0.00005)

3

Replacement Method: zero for all <DL

- **This is the impact of substituting an artificial value (in this case zero) into a data set: the mean is far too low and the variance (square of the standard deviation) far too large**

	Complete Data	Substitute zero
Mean	**1.108**	**0.747**
Variance	**0.117**	**0.505**

4

Replacement Method: DL for all <DL

- This is the impact of substituting an artificial value (in this case the DL) into a data set: the mean is far too high and the variance (square of the standard deviation) far too small

	Complete Data	Substitute DL
Mean	1.108	1.422
Variance	0.117	0.099

5

Replacement Method: DL/2

- This is the impact of substituting artificial values into a data set: the mean is too low and the variance (square of the standard deviation) too large

	Complete Data	Substitute DL/2
Mean	1.108	0.972
Variance	0.117	0.302

6

Cohen's Method: Assumptions

- This method requires that the data without the nondetects be Normally distributed (non-detect data being the lower tail of the Normal but not recorded of course)
- This method assumes that there is only one detection limit in the data set
- This is known as a Maximum Likelihood Method because of the assumption of Normality

7

Cohen's Method: Setup

- Let $X_1, X_2, ..., X_m, ..., X_n$ represent all the n data values ranked from largest to smallest
- The first "m" values represent the data values above the detection limit (DL), and the remaining "n-m" data points are those below the detection limit

8

Cohen's Method: Calculations

- Compute the sample mean and the sample variance from only the "m" above detection data values
- Obviously the mean will be too large as we have ignored the small undetected values, and the variance too small as we have effectively compacted together the data

$\bar{X}_d$ is the sample mean

s_d is the sample standard deviation

m is the number of detected values

n is the total number of values

9

Cohen's Method: Obtaining values

- The mean will be lowered and the variance enlarged through the use of factors:

$$h = \frac{n - m}{n} \qquad \gamma = \frac{s_d^2}{\left(\bar{X}_d - DL\right)^2}$$

- These are used together in a set of tables (EPA/QA/G-9S, Table A-11) to find the value:

$$\hat{\lambda}(\gamma, h)$$

10

Cohen's Method: Adjusting

- Estimate the corrected sample mean and corrected sample variance to account for the data below the detection limit:

$$\overline{X} = \overline{X}_d - \hat{\lambda}\left(\overline{X}_d - DL\right)$$

$$s^2 = s_d^2 + \hat{\lambda}\left(\overline{X}_d - DL\right)^2$$

11

Cohen's Method: Arsenic Example

- Let $X_1, X_2, ..., X_m, ..., X_n$ represent all the n data values ranked from largest to smallest
  1.752, 1.563, 1.498, 1.477, 1.418, 1.358, 1.327, 1.289, 1.148, 1.060, 1.045, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000
- The first "m" values represent the data values above the detection limit (DL), and the remaining "n-m" data points are those below the detection limit
  n = 20, m = 11, n-m = 9
- Compute the sample mean and the sample variance from only the “m” above detection data values
  Sample Mean = 1.358 Sample Variance = 0.0524

12

Cohen's Method: Obtaining values for Arsenic example

- The first factor is then 11/20 = 0.55
- The second factor is then $0.0524/(1.358 - 1.000)^2 = 0.409$
- The third factor uses these two factors and Table A-11 = 1.113

13

Table A-11

	h											
γ	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.80	.90
.00	.31862	.4021	.4941	.5961	.7096	.8388	.9808	1.145	1.336	1.561	2.176	3.283
.05	.32793	.4130	.5066	.6101	.7252	.8540	.9994	1.166	1.358	1.585	2.203	3.314
.10	.33662	.4233	.5184	.6234	.7400	.8703	1.017	1.185	1.379	1.608	2.229	3.345
.15	.34480	.4330	.5296	.6361	.7542	.8860	1.035	1.204	1.400	1.630	2.255	3.376
.20	.35255	.4422	.5403	.6483	.7673	.9012	1.051	1.222	1.419	1.651	2.280	3.405
.25	.35993	.4510	.5506	.6600	.7810	.9158	1.067	1.240	1.439	1.672	2.305	3.435
.30	.36700	.4595	.5604	.6713	.7937	.9300	1.083	1.257	1.457	1.693	2.329	3.464
.35	.37379	.4676	.5699	.6821	.8060	.9437	1.098	1.274	1.475	1.713	2.353	3.492
.40	.38033	.4735	.5791	.6927	.8179	.9570	1.113	1.290	1.494	1.732	2.376	3.520
.45	.38665	.4831	.5880	.7029	.8295	.9700	1.127	1.306	1.511	1.751	2.399	3.547
.50	.39276	.4904	.5967	.7129	.8408	.9826	1.141	1.321	1.528	1.770	2.421	3.575
.55	.39679	.4976	.6061	.7225	.8517	.9950	1.155	1.337	1.545	1.788	2.443	3.601
.60	.40447	.5045	.6133	.7320	.8625	1.007	1.169	1.351	1.561	1.806	2.465	3.628
.65	.41008	.5114	.6213	.7412	.8729	1.019	1.182	1.368	1.577	1.824	2.486	3.654
.70	.41555	.5180	.6291	.7502	.8832	1.030	1.195	1.380	1.593	1.841	2.507	3.679
.75	.42090	.5245	.6367	.7590	.8932	1.042	1.207	1.394	1.608	1.851	2.528	3.705
.80	.42612	.5308	.6441	.7676	.9031	1.053	1.220	1.408	1.624	1.875	2.548	3.730
.85	.43122	.5370	.6515	.7781	.9127	1.064	1.232	1.422	1.639	1.892	2.568	3.754
.90	.43622	.5430	.6586	.7844	.9222	1.074	1.244	1.435	1.653	1.908	2.588	3.779
.95	.44112	.5490	.6656	.7925	.9314	1.085	1.255	1.448	1.668	1.924	2.607	3.803
1.00	.44592	.5548	.6724	.8005	.9406	1.095	1.287	1.461	1.882	1.940	2.626	3.827

14

Cohen's Method: Adjusting for the Arsenic Example

- **Estimate the corrected sample mean and corrected sample variance to account for the data below the detection limit:**

$$\overline{X} = \overline{X}_d - \hat{\lambda}\left(\overline{X}_d - DL\right)$$

= 1.358 – 1.113(1.358 – 1) = 0.960

$$s^2 = s_d^2 + \hat{\lambda}\left(\overline{X}_d - DL\right)^2$$

= 0.0524 + 1.113(1.358 – 1)^2 = 0.195

15

Cohen's Maximum Likelihood Method

- **This is the impact of assuming Normality substituting estimated values into a data set: the mean is a little too low and the variance (square of the standard deviation) a little too large. Maximum likelihood estimation needs a fair amount of data to make the initial estimates.**

	Complete Data	Cohen's method
Mean	**1.108**	**0.960**
Variance	**0.117**	**0.195**

16

Regression on Order Statistics Method

- ROS uses a Normal Probability Plot with the Y-values being the actual values observed, the X-values being the inverse standard normal cumulative density function.
- Imagine a Normal curve superimposed on the data and then see how well the data fit the curve.
- Regression on Order Statistics essentially bends the Normal curve back to a straight line (X-axis) that has the spacing between x-values unequal but related to a Normal curve.
- Ordinary least squares regression is used to find the intercept (the mean) and slope of the line (the standard deviation.

17

Regression on Order Statistics Method

- This is the impact of assuming Normality and using Regression on Order Statistics; intercept being 1.197 and the slope of the line being 0.219. The mean is fairly well estimated but the variance (square of the standard deviation) too small.

	Complete Data	ROS method
Mean	1.108	1.197
Variance	0.117	0.048

18

Kaplan-Meier Method: Assumptions

- This method does not require knowing the distribution of the data
- This method can accommodate multiple different detection limits in the data set
- This is known as the flip version of censored survival data analysis

19

Arsenic in Soil Samples

- The following data values have been recorded and submitted for analysis and comment (note: there are now 2 detection limits; 0.800 and 1.000):

1.752	0.944	1.418	1.477	0.897
<0.800 (0.777)	1.289	1.498	0.818	<0.800 (0.622)
1.327	0.919	1.060	<0.800 (0.455)	1.358
0.977	0.622	<1.000 (0.806)	1.563	1.148

(Actually the method is accurate to 3 decimal places and can measure reliably down to its own limit of 0.00005)

20

Kaplan-Meier Method: The set up

- This method is an adaptation of censored survival data analysis which investigates the time to failure or death.
- To make the method applicable to detection limit problems we must chose a number well clear of the real data set and then "flip" the data around this number
- For the Arsenic example the maximum was 1.752 and so we can chose 2 (or 3 or 4, it makes no difference) and our flip point.
- 1.752 when flipped is 0.248, 1.563 becomes 0.437, and the smallest recorded value, 0.622 becomes 1.378.

21

Kaplan-Meier Method: The method

- This method will find a specific probability (denoted as g_i) for each X_i (the flipped values) using an "Incremental Survival Probability" (actually through use of a table you have to construct).
- The "g_i" and "X_i" will be put together to make estimates of the mean and variance:

$$\text{Mean} = \sum g_i X_i$$

$$\text{Variance} = \sum g_i X_i^2 - (\text{Mean})^2$$

- The Mean is then flipped back to the original scale but the variance is left as it is.

22

Kaplan-Meier Method: The calculation

Data	Flip on 2	rank	b = n-r+1	d	p=(b-d)/b	S	g
1.752	0.248	1	20	1	19/20	0.9500	0.0500
1.563	0.437	2	19	1	18/19	0.9000	0.0500
1.498	0.502	3	18	1	17/18	0.8500	0.0500
1.477	0.523	4	17	1	16/17	0.8000	0.0500
1.418	0.582	5	16	1	15/16	0.7500	0.0500
1.358	0.642	6	15	1	14/15	0.7000	0.0500
1.327	0.673	7	14	1	13/14	0.6500	0.0500
1.289	0.711	8	13	1	12/13	0.6000	0.0500
1.148	0.852	9	12	1	11/12	0.5500	0.0500
1.060	0.940	10	11	1	10/11	0.5000	0.0500
1.045	0.955	11	10	1	9/11	0.4500	0.0500
0.977	1.023	13	8	1	8/9	0.3938	0.05625
0.944	1.056	14	7	1	7/8	0.3375	0.05625
0.919	1.081	15	6	1	6/7	0.2813	0.05625
0.897	1.103	16	5	1	5/6	0.2250	0.05625
0.818	1.182	17	4	1	4/5	0.1688	0.05625
<0.800	>1.200	18	3	3	0	0	0.16875

23

Kaplan-Meier Method: Explained

Col 1: The actual data values (non-detects indicated by a dashed line)
Col 2: The "flipped data" = 2 minus the actual value
Col 3: Rank order (the missing ranks belong to non-detects)
Col 4: b = n–r+1 where n= total (20), r = rank
Col 5: d = number of observations for this value (1 in this case)
Col 6: p = (b - d)/b
Col 7: S = The S from the previous row multiplied by the p for the current row (starts at 1.0000)
e.g. 10th data value: S = 0.5500 x 10/11 = 0.5000
Col 8: g = The S from the previous row minus the S for the current row (starts at 1.0000)
e.g. 10th data value: g = 0.5000 – 0.4500 = 0.0500

24

Kaplan-Meier Method: Mean & Variance

Mean = $\sum g_i X_i$ and Variance = $\sum g_i X_i^2$ - $(\text{Mean})^2$

The X_is are the flipped values and the g_is come from the table

Mean = 0.05x0.248 + …+ 0.16875x1.200 = 0.8620

Variance = $0.05 \times 0.248^2 + \ldots + 0.16875 \times 1.200^2 - 0.8620^2 = 0.085$

The true Mean is then 2 – 0.8620 = 1.138 and the variance 0.085

25

Kaplan-Meier Method

- This shows the non-parametric approach to dealing with non-detects; the mean is just a little too low and the variance (square of the standard deviation) just a little too large – both are very close to the true values.

	Complete Data	Kaplan-Meier
Mean	1.108	1.138
Variance	0.117	0.085

26

Comparison of Methods

	True	Zero	DL	½ DL	Cohen	ROS	K-M
Mean	1.108	0.747	1.422	0.972	0.960	1.197	1.138
Var	0.117	0.505	0.099	0.302	0.195	0.048	0.085

- ½ DL: Not too good but very easy to find the estimates
- Cohen/ROS: Fairly good considering the high non-detect rate (45%) but the methods need work in calculating estimates
- Kaplan-Meier: Very good, with a positive point that the method can account for multiple detection limits, but considerable work is needed to calculate the estimates

27

Summary

Percentage of Nondetects	Statistical Analysis Method
<~15%	Replace non-detects with DL/2
15% - 50%	Cohen's Maximum Likelihood Method Kaplan-Meier's Nonparametric Method
>50%	Problem: No method preferred

28

					h					
;	.30	.35	.40	.45	.50	.55	.60	.65	.70	.8
52	.4021	.4941	.5961	.7096	.8388	.9808	1.145	1.336	1.561	2.1
93	.4130	.5066	.6101	.7252	.8540	.9994	1.166	1.358	1.585	2.2
52	.4233	.5184	.6234	.7400	.8703	1.017	1.185	1.379	1.608	2.2
80	.4330	.5296	.6361	.7542	.8860	1.035	1.204	1.400	1.630	2.2
55	.4422	.5403	.6483	.7673	.9012	1.051	1.222	1.419	1.651	2.2
93	.4510	.5506	.6600	.7810	.9158	1.067	1.240	1.439	1.672	2.3
00	.4595	.5604	.6713	.7937	.9300	1.083	1.257	1.457	1.693	2.3
79	.4676	.5699	.6821	.8060	.9437	1.098	1.274	1.475	1.713	2.3
33	.4735	.5791	.6927	.8179	.9570	1.113	1.290	1.494	1.732	2.3
65	.4831	.5880	.7029	.8295	.9700	1.127	1.306	1.511	1.751	2.3
76	.4904	.5967	.7129	.8408	.9826	1.141	1.321	1.528	1.770	2.4
79	.4976	.6061	.7225	.8517	.9950	1.155	1.337	1.545	1.788	2.4
47	.5045	.6133	.7320	.8625	1.007	1.169	1.351	1.561	1.806	2.4
08	.5114	.6213	.7412	.8729	1.019	1.182	1.368	1.577	1.824	2.4
55	.5180	.6291	.7502	.8832	1.030	1.195	1.380	1.593	1.841	2.5
90	.5245	.6367	.7590	.8932	1.042	1.207	1.394	1.608	1.851	2.5
12	.5308	.6441	.7676	.9031	1.053	1.220	1.408	1.624	1.875	2.5
22	.5370	.6515	.7781	.9127	1.064	1.232	1.422	1.639	1.892	2.5
22	.5430	.6586	.7844	.9222	1.074	1.244	1.435	1.653	1.908	2.5
12	.5490	.6656	.7925	.9314	1.085	1.255	1.448	1.668	1.924	2.6
92	.5548	.6724	.8005	.9406	1.095	1.287	1.461	1.882	1.940	2.6

Kaplan-Meier Method: The set up

- First chose a number ("flip point") well clear of the real data set and then "flip" the data around this number; these are now called the flipped values
- This method will find a specific probability (denoted as g_i) for each X_i (the flipped values) using a table
- The "g_i" and "X_i" will then be put together:

$$\text{Mean} = \sum g_i X_i$$

$$\text{Variance} = \sum g_i X_i^2 - (\text{Mean})^2$$

The Mean is then flipped back to the original scale and the variance is left as it has just been calculated.

1

Kaplan-Meier Method: The calculation

Data	Flipped	rank	b = n-r+1	d	p=(b-d)/b	S	g

2

Kaplan-Meier Method: Explained

Col 1: The actual data values
Col 2: The "flipped data" = Flip point minus the actual value
Col 3: Rank order (the missing ranks belong to non-detects)
Col 4: b = n–r+1 where n= total, r = rank
Col 5: d = number of observations for this value
Col 6: p = (b - d)/b
Col 7: S = The S from the previous row multiplied by the p for the current row (starts at 1.0000)
Col 8: g = The S from the previous row minus the S for the current row (starts at 1.0000)

3

Actual data :Tetrachloroethelene

Data value concentration units are $\mu g/m^3$

0.81, <0.68, <0.68, 0.95, <0.68, <0.68, <0.68
<0.68, 9.29, 1.90, 0.88, 2.98, 0.75, 5.97

Note that the total number of observations is 14, of which 6 are below detection (42.9%)

Step 1: Rearrange the data values largest to smallest: 9.29, 5.97, 2.98, 1.90, 0.95, 0.88, 0.81, 0.75, and then <0.68 six times

Step 2: As the largest value is 9.29 let us choose 10.00 to be the flip point

4

Kaplan-Meier Method: Mean & Variance

X_i	g_i	g_iX_i	X_i^2	$g_iX_i^2$
0.71				
4.03				
7.02				
8.10				
9.05				
9.12				
9.19				
9.25				
>9.32				
---	---	Σg_iX_i	---	$\Sigma g_iX_i^2$

5

Kaplan-Meier Method: Mean & Variance

X_i	g_i	g_iX_i	X_i^2	$g_iX_i^2$
0.71			0.5041	
4.03			16.2409	
7.02			49.2804	
8.10			65.6100	
9.05			81.9025	
9.12			83.1744	
9.19			84.4561	
9.25			85.5625	
>9.32			86.8624	
---	---		---	

6

Kaplan-Meier Method: Mean & Variance

Mean = $\sum g_i X_i$ and Variance = $\sum g_i X_i^2$ - (Mean)^2

The X_is are the flipped values and the g_is come from the table

Mean =
Variance =

The true Mean is then Flip point minus the Mean from above, Variance is the same as just calculated

7

Conclusions

Actual data :Tetrachloroethelene

Data value concentration units are µg/m^3
Cancer 10^{-5} risk concentration is 1.7µg/m^3

0.81, <0.68, <0.68, 0.95, <0.68, <0.68, <0.68
<0.68, 9.29, 1.90, 0.88, 2.98, 0.75, 5.97

Note that the total number of observations is 14, of which 6 are below detection (42.9%)

Choices:
a) Replace <0.68 by 0.34 (DL/2)
b) Adjust using Cohen's Method
c) Use Kaplan-Meier Method

2

Comparison of methods

a) Half Detection Level substitution
Estimated mean = 1.826
Estimated variance = 7.027
We know its biased but unknown which way

b) Cohen's Maximum Likelihood Method
Estimated mean = 0.941
Estimated variance = 14.256
We know this could be biased if not Normal data

c) Kaplan-Meier's Nonparametric Method
Estimated mean = 1.974
Estimated variance = 6.129
We know this is slightly biased a little too high

3

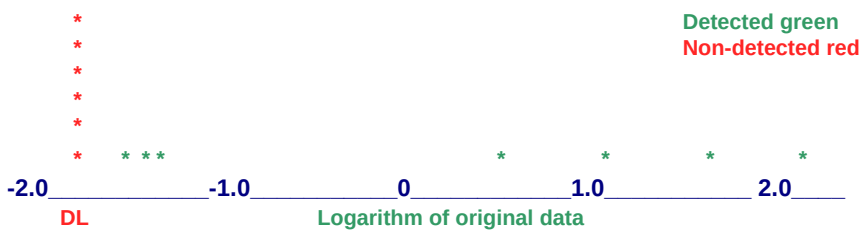
Another look at the data (1)

```
*                                                                  Detected green
*                                                                  Non-detected red
*
*
*
* ****      *      *                  *                      *
0____ 1_____2_____3_____4_____5_____6_____7_____8_____9____10
DL                                Original data
```

Could there be some outliers in the data?
Could the <DL values be uniformly spread out below the DL?
Could the data be (or assumed to be) Normal/Lognormal?
Could this be something strange?

4

Another look at the data (2)



Could there be some outliers in the data?
Could the <DL values be uniformly spread out below the DL?
Could the data be (or assumed to be) Normal/Lognormal?
Could this be something strange?

5

What to do when non-detects > 50%

The absence of real values stresses the theoretical solutions and bias starts to become a serious problem

Some possibilities:

- **Substitute 1/2 Detection Level**
- **Cohen's Maximum Likelihood Assuming Normality**
- **Kaplan-Meier (Adjusted)**
- **Maximum Likelihood Assuming Lognormality**
- **Regression on Order Statistics**

but these take great effort and all have significant problems depending on number of samples, distribution, and DL.

6

But which one shall I use?

One way to discuss the performance of rival methods is to use simulation techniques. A small sample is taken from a known distribution of data, censored at some desired level, and estimates calculated by some method. This is repeated thousands of times and the results compared to see how each well each method did in comparison with the known distribution.

Another method is to test the competing methods using data sets that are known relatively accurately. These data sets are specially chosen to be representative of the type of data sets encountered in practice.

7

Has anyone done this?

For simulation studies there are at least 12 investigations and these are summarized in *Nondetects and Data Analysis,* 2005 by Dennis R. Helsel (Chapter 6).

For comparison studies the most recent is *Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets*, 2008 by Ronald C. Antweiler and Howard E. Taylor, ES&T, 42, No.10.

They do not completely agree on the best methods to use…

8

Simulation (S) versus Comparison (C)

Type of data used:
S: Known distribution datasets
C: ***Actual environmental datasets***

Influence of sample size on results:
S: Seemed to be a change around n = 50
C: Made only minor differences

Amount of censoring where all methods are poor
S: Between 60% - 80%
C: Roughly 70%

9

Simulation studies

Helsel's recommendations:

Percent Censored	< 50 observations	> 50 observations
<50% non-detects	Kaplan-Meier	Kaplan-Meier
50% - 80% non-detects	Maximum Likelihood Reg. on Order Stats	Maximum Likelihood
>80% non-detects	Report only % above a meaningful threshold	Report only high sample percentiles (95th, 99th)

10

Comparison studies

Antweiler and Taylor's recommendations:

Parameter estimated	<70% censored data	>70% censored data
Mean	DL/2: quite good Kaplan-Meier: very good	DL/2: fair Kaplan-Meier: quite good Reg on Ord Stat: quite good
Standard deviation	DL/2: good Kaplan-Meier: very good	DL/2: good Kaplan-Meier: good Reg on Ord Stat: quite good

11

What to do, what to do…

The choice of method to use should be related to the intended use of the data. Some (informal) recommendations:

	Fairly small sample	Fairly large sample	Quite large sample
Exploratory use	DL/2 (provided only a few <DL)	DL/2 (provided less than approx 15% <DL)	Cohen (if Normal, Kaplan-Meier otherwise)
Publication use	Kaplan-Meier	Kaplan-Meier Cohen (if apprx Normal)	Cohen (if Normal, Kaplan-Meier otherwise)
Regulatory use	Kaplan-Meier	Kaplan-Meier	Kaplan-Meier

12

Visual Sample Plan Demonstration

What is VSP?

- Data Quality Objectives based systematic planning software to determine the number and location of environmental samples
- Provides statistically defensible decisions
- Has the capability to perform statistical and data quality assessments
- Sponsored by different government agencies: EPA, DoD, DOE, CDC, UKAWE

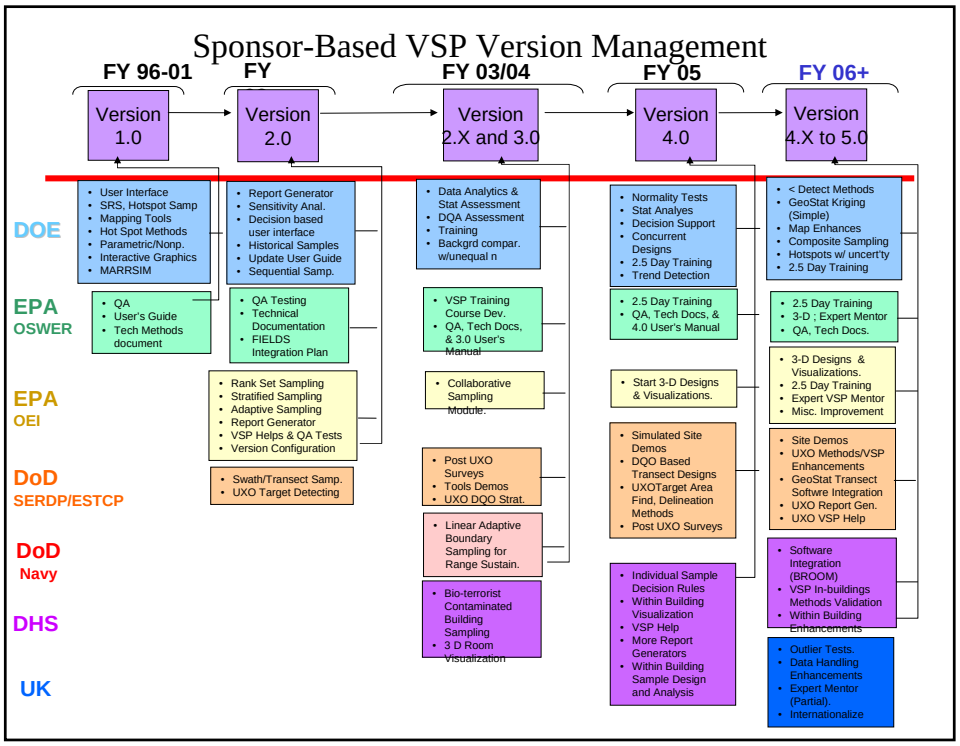
Available free at http://vsp.pnl.gov

2

VSP Sponsors

- U.S. Environmental Protection Agency (EPA)
  - Office of Solid Waste and Emergency Response
  - Office of Environmental Information
- U.S. Dept. of Energy (DOE)
  - HS, LM, EM
- U.S. Dept. of Defense (DoD)
  - SERDP/ESTCP
- U.S. Dept. of Homeland Security (DHS)
  - Technical Support Working Group (TSWG)
  - Chemical and Biological Countermeasures
- U.S. Centers for Disease Control (CDC/NIOSH)
- U.K. Atomic Weapons Establishment (UKAWE)

3

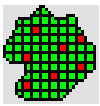


Multi-Agency Leveraging

- Each Agency has their own focus and needs
- Some capability requirements are shared
- No one agency can afford to support VSP development but jointly we can
- The total is greater than the sum of its parts!

5

Visual Sample Plan



- Designed for the non-statistician
- Has visual, graphical, user-friendly interface
- Provides summary statistics of data sets
- Performs statistical and graphical tests
- Focused on hazardous waste sites, surface contamination, UXO contamination and contamination of rooms

6

VSP Design Features

- Various statistical sampling options:
  - Simple Random Sampling
  - Systematic Grid Sampling
  - Sequential Sampling
  - Collaborative Sampling
  - Stratified Sampling
  - Rank-Set Sampling
  - Adaptive Cluster Sampling
  - Continuous Transect Sampling
  - Judgmental Sampling

7

VSP Analysis Features

- Interactive diagnostic graphics for tradeoff analyses
- Visual features with locations displayed
- Normal and non-Normal distribution methods
- Sensitivity analysis features
- Automatic report generator
- Online help, user manual, technical documents
- Documented VSP portfolio QA plan

8

Visual Sample Plan and Data Containing Non-detects

Actual data :Tetrachloroethelene

Data value concentration units are µg/m^3

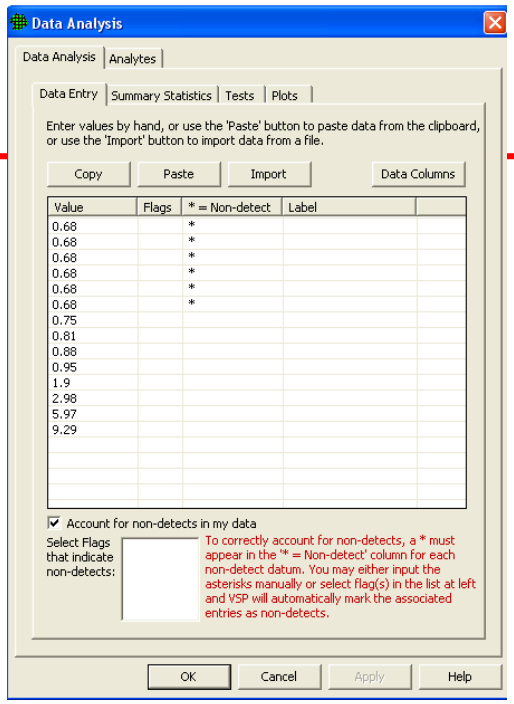
0.81, <0.68, <0.68, 0.95, <0.68, <0.68, <0.68
<0.68, 9.29, 1.90, 0.88, 2.98, 0.75, 5.97

Note that the total number of observations is 14, of which 6 are below detection (42.9%)

Rearrange the data values smallest to largest for clarity:

<0.68, <0.68, <0.68, <0.68, <0.68, <0.68
0.75, 0.81, 0.88, 0.95, 1.90, 2.98, 5.97, 9.29

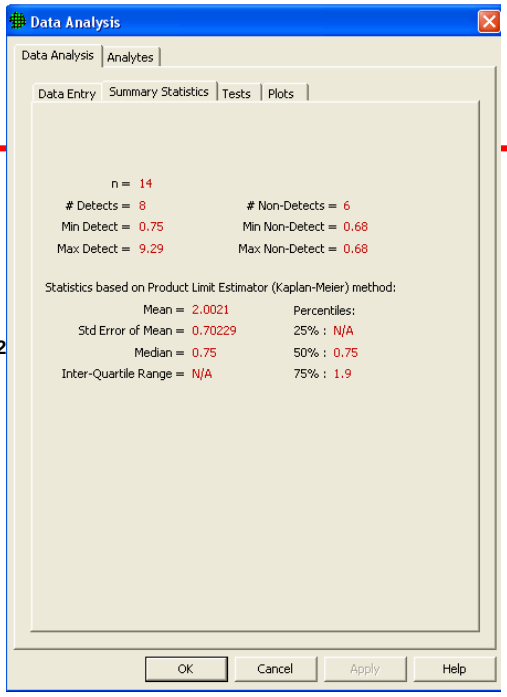
2



3

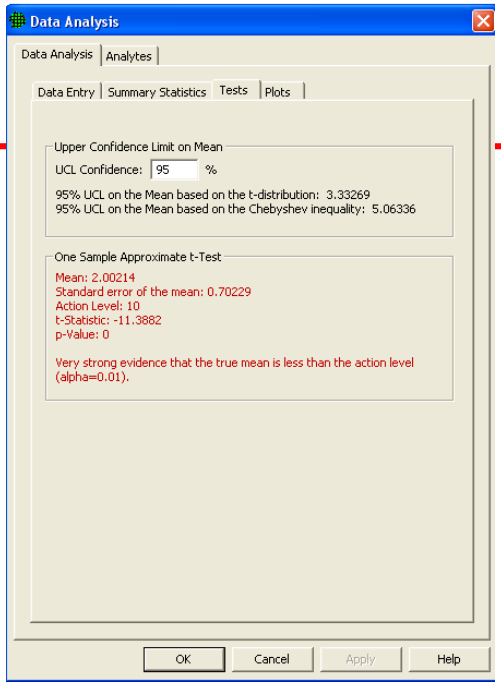
Variance =

n x (SEmean)^2



14 x 0.70229^2
= 6.90496

4



5

Kaplan- Meier Method: Conclusion

From the hand computed method it was found:
Estimated mean = 1.9721
Estimated variance = 6.11819

From Visual Sample Plan the results were:
Estimated mean = 2.0021
Estimated variance = 6.90496

Why the discrepancy?
Hand computed method used the fact the detection limit was known (0.68); Visual Sample Plan did not do this. VSP used the fact that that the detection limit had to be less than the smallest observed value (0.81) and so used 0.81 in its calculations.

6

Visual Sample Plan and Data Containing Multiple Non-detects

Arsenic in Soil Samples

Data value concentration units are µg/m^3

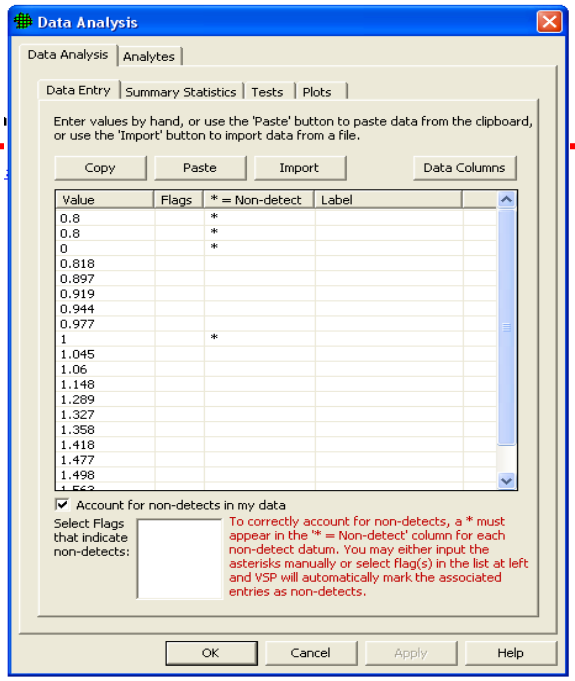
1.752, 0.944, 1.418, 1.477, 0.897, <0.800, 1.289, 1.498, 0.818, <0.800, 1.327, 0.919, 1.060, <0.800, 1.358, 0.977, 1.045, <1.000, 1.563, 1.148

Note that the total number of observations is 20, of which 3 are below detection 0.800, 1 is below detection 1.000

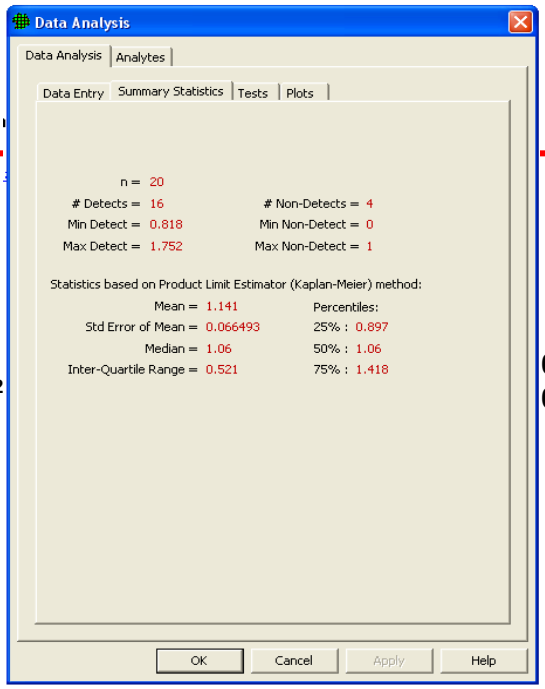
Rearrange the data values smallest to largest for clarity:

<0.800, <0.800, <0.800, 0.818, 0.897, 0.919, 0.944, 0.977, <1.000, 1.045, 1.060, 1.148, 1.289, 1.327, 1.358, 1.418, 1.477, 1.498, 1.563, 1.752

2



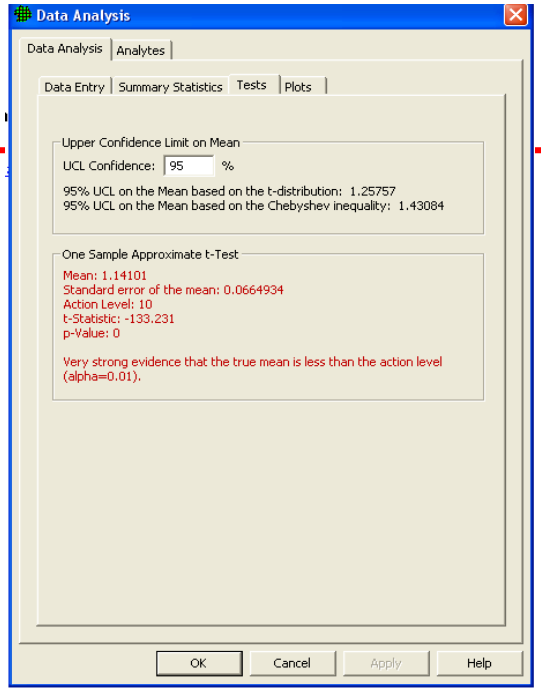
3



Variance = n x (SEmean)^2

0.085 = 20 x 0.066493^2

4



5

Kaplan- Meier Method: Conclusion

From the hand computed method it was found:
Estimated mean = 1.138
Estimated variance = 0.085

From Visual Sample Plan the results were:
Estimated mean = 1.141
Estimated variance = 0.088

Why the discrepancy?
Hand computed method used the fact the two detection limits were known (0.8 & 1.0). VSP used the fact that the upper detection limit had to be less than the smallest value above it (1.045) and the lower detection limit had to be less than the smallest observed value (0.818) and so used these in its calculations.

6

Visual Sample Plan and the Analysis of Data

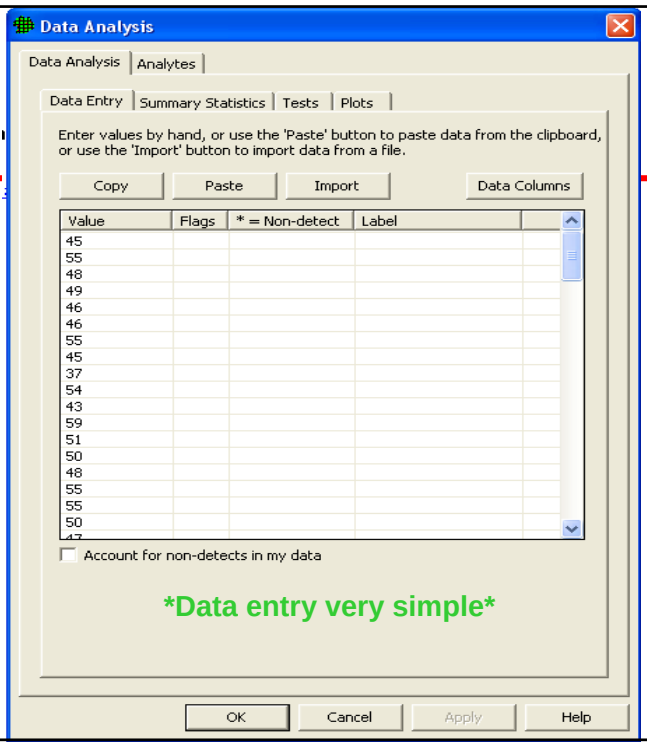
Comstock Air Monitoring Station

Ozone ppb

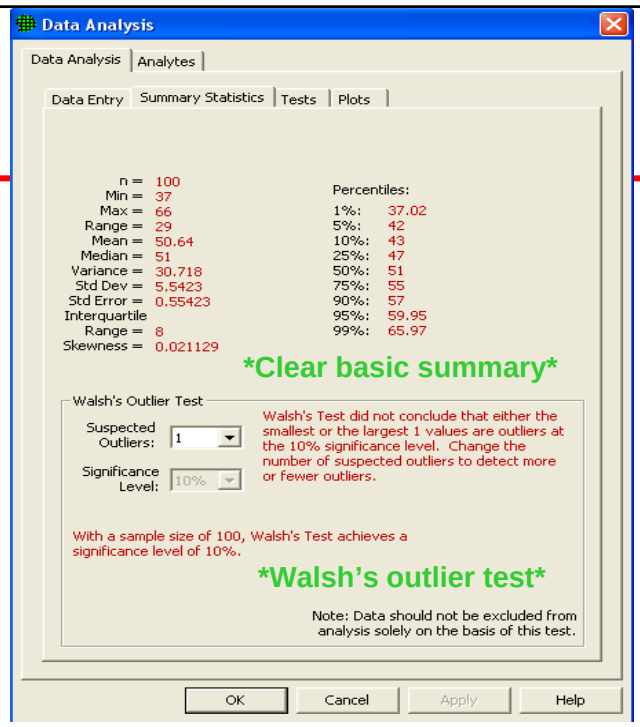
45	48	48	52	59
55	51	47	50	66
48	49	48	48	61
49	54	40	46	58
46	47	55	45	54
46	42	50	49	49
55	43	53	54	47
45	40	56	55	44
37	57	48	49	43
54	55	57	52	48
43	60	56	52	51
59	51	50	52	57
51	42	46	55	52
50	44	48	55	52
48	55	49	42	55
55	39	53	47	51
55	43	57	51	53
50	46	63	52	55
47	48	58	55	60
43	47	53	54	57

100 consecutive days, Summer 2005

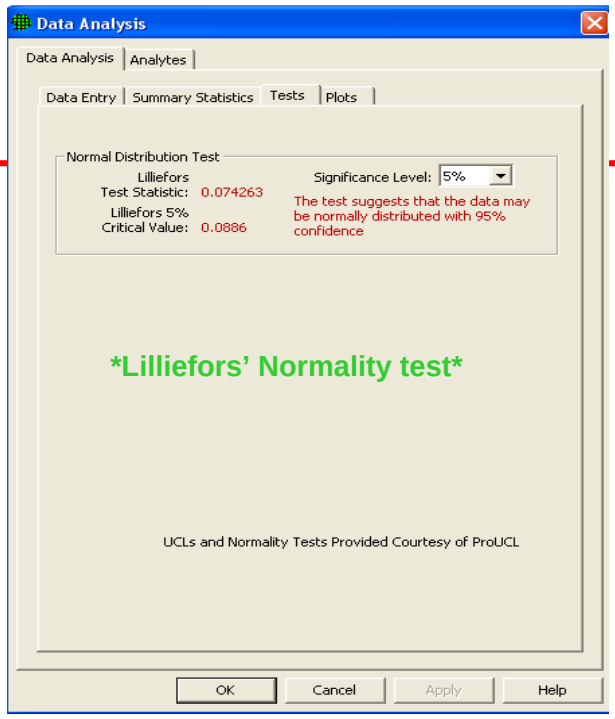
2


Data Analysis
Data Analysis
Analytes
Data Entry
Summary Statistics
Tests
Plots
Enter values by hand, or use the 'Paste' button to paste data from the clipbaord, or use the 'Import' button to import data from a file.
Copy
Paste
Import
Data Columns
Value
Flags
* = Non-detect
Label
45
55
48
49
46
46
55
45
37
54
43
59
51
50
48
55
55
50
Account for non-detects in my data
Data entry very simple
OK
Cancel
Apply
Help

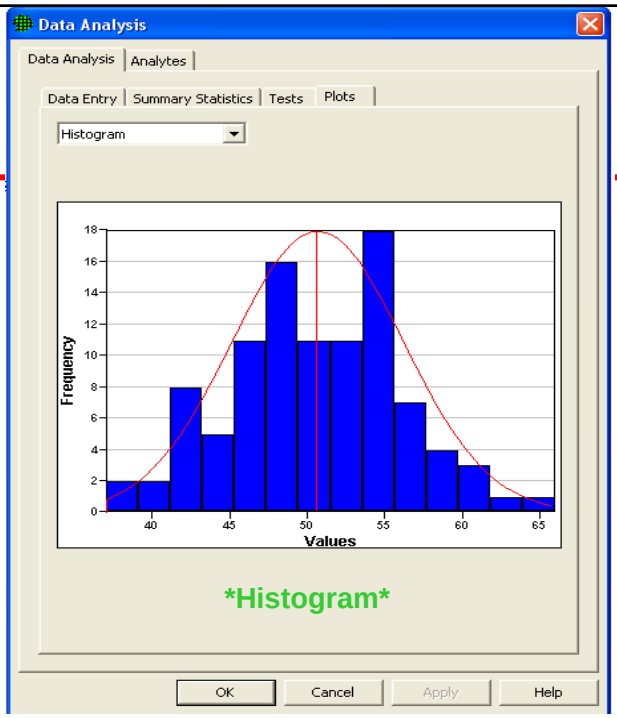
3


Data Analysis
Data Analysis
Analytes
Data Entry
Summary Statistics
Tests
Plots
n = 100
Min = 37
Max = 66
Range = 29
Mean = 50.64
Median = 51
Variance = 30.718
Std Dev = 5.5423
Std Error = 0.55423
Interquartile Range = 8
Skewness = 0.021129
Percentiles:
1%: 37.02
5%: 42
10%: 43
25%: 47
50%: 51
75%: 55
90%: 57
95%: 59.95
99%: 65.97
Clear basic summary
Walsh's Outlier Test
Suspected Outliers: 1
Significance Level: 10%
Walsh's Test did not conclude that either the smallest or the largest 1 values are outliers at the 10% significance level. Change the number of suspected outliers to detect more or fewer outliers.
With a sample size of 100, Walsh's Test achieves a significance level of 10%.
Walsh's outlier test
Note: Data should not be excluded from analysis solely on the basis of this test.
OK
Cancel
Apply
Help

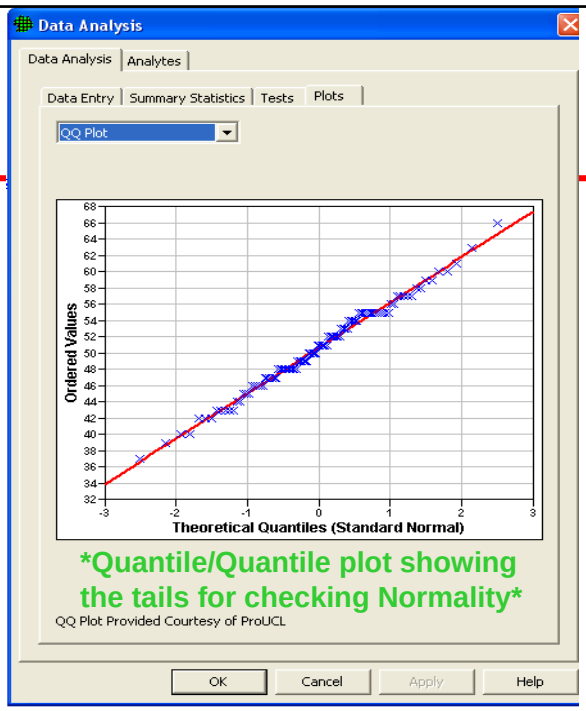
4


Data Analysis
Data Analysis
Analytes
Data Entry
Summary Statistics
Tests
Plots
Normal Distribution Test
Lilliefors
Test Statistic: 0.074263
Lilliefors 5%
Critical Value: 0.0886
Significance Level: 5%
The test suggests that the data may be normally distributed with 95% confidence
Lilliefors' Normality test
UCLs and Normality Tests Provided Courtesy of ProUCL
OK
Cancel
Apply
Help

5


Data Analysis
Data Analysis
Analytes
Data Entry
Summary Statistics
Tests
Plots
Histogram
Frequency
Values
Histogram
OK
Cancel
Apply
Help

6



Quantile/Quantile plot showing the tails for checking Normality

QQ Plot Provided Courtesy of ProUCL

7

VSP Conclusions

- VSP is primarily a planning tool
- VSP is also an analysis tool
- Extensively peer-reviewed
- Has EPA documents as its basis
  - DQO Planning Process (QA/G-4)
  - Sampling Designs (QA/G-5S)
  - DQA Statistical Tools (QA/G-9S)
- Free: http//:vsp.pnl.gov

8