EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Robust Statistical Approach to Analysis of Data Containing Non-detects

Babu Nott, Christine Lee – EPRI

Dennis Helsel – Practical Stats

National Environmental Monitoring Conference

August 4-8, 2014

Washington DC

Presentation Outline

- Background
- Simple substitution for the detection limit produces poor estimates
- Alternative methods are available and common in other areas of science
- We demonstrate use of the Kaplan-Meier method to selected data on emissions from electric power plants

Background

- Issue: real-world environmental data may contain measured (quantified) values as well as non-detects
- Simple substitution is used for less than detection/ quantitation values in environmental rulemaking
  - 2012: Mercury and Air Toxics Standards (MATS)
  - 2013: Proposed Effluent Limitation Guidelines (ELG)
- Substituting values for non-detects is arbitrary and unnecessary
- Alternatives to substitution methods have been successfully used in other fields – e.g., health sciences, risk analysis,…

EPA's 2010 ICR for EGU HAPs Emissions

- MATS relied on EPA's 2010 Information Collection Request (ICR) on Hazardous Air Pollutants (HAPs) from Electric Generating Units (EGUs)
  - Stack gas emissions from coal and oil-fired power plants
  - HAPs: Metals including mercury, acid gases, organics
  - Data consisted of measurements both above and below detection levels
  - EPA used zero, ½ DL or DL in place of non-detects at different points in their data analyses
- Why is this a problem?
  - Substitution produces poor estimates of variability, which directly affects computation of regulatory limits

Nondetects

- Called “censored data” in statistics – known only as above or below a threshold
- contain much of the information present in detected values

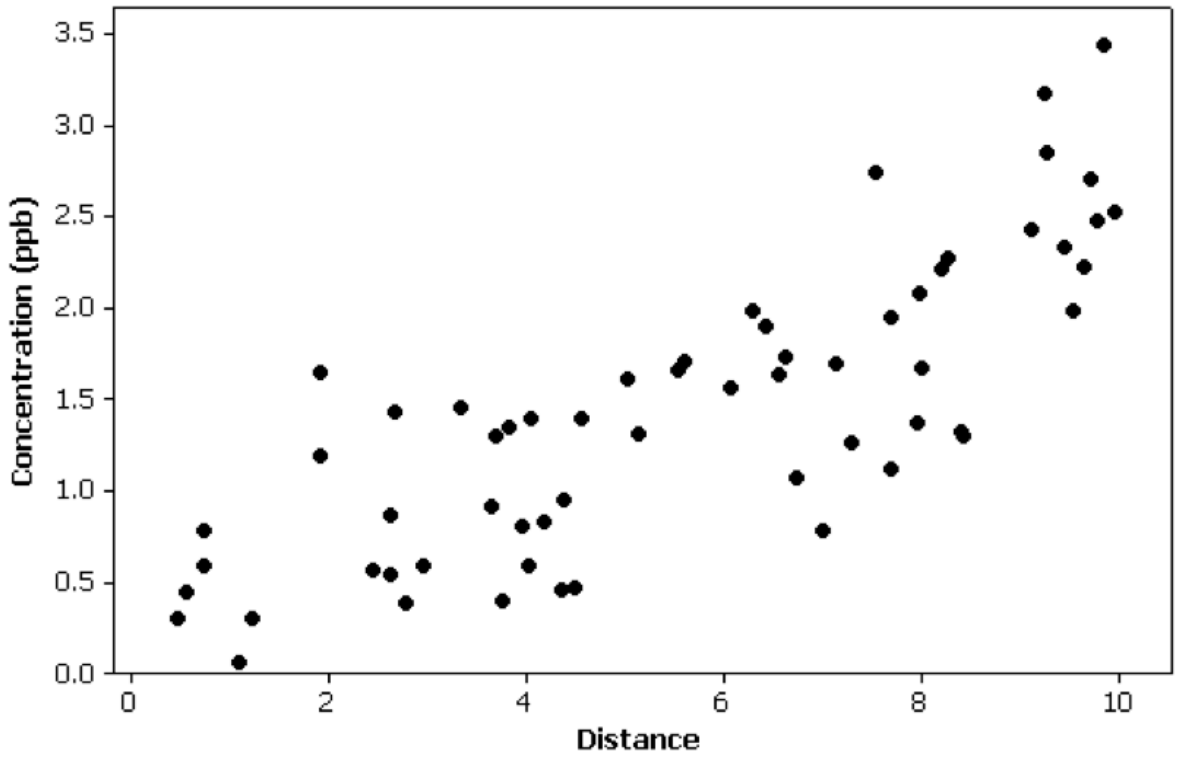
Methods for Censored Data

- Have been used in medical and industrial statistics since the 1950s
- There they usually are “greater-thans”, i.e. >10 , otherwise the issues are the same
- This area of statistics is called "Survival analysis" or "Reliability analysis"
- These methods are used in drug trials, testing equipment and processes, occupational health, astronomy, and in many other fields
- For more detail, see the text book *Statistics for Censored Environmental Data* by Helsel (2012)

What's wrong with substitution?

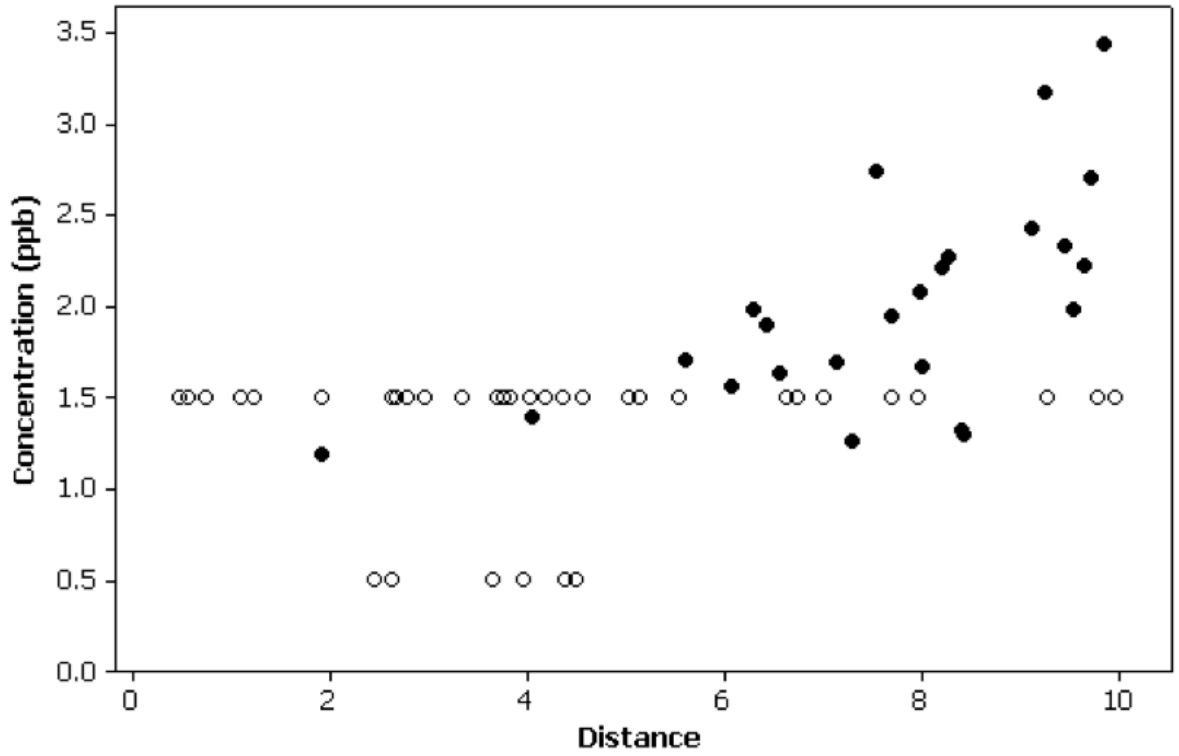
- Produces invasive data alien to the concentrations actually in samples
- Adds a pattern that likely was not in the data. Substitution is NOT neutral
- Could result in finding “No Difference Between Groups” or “No Trend” when there is one, and vice versa

Example 1: Regression



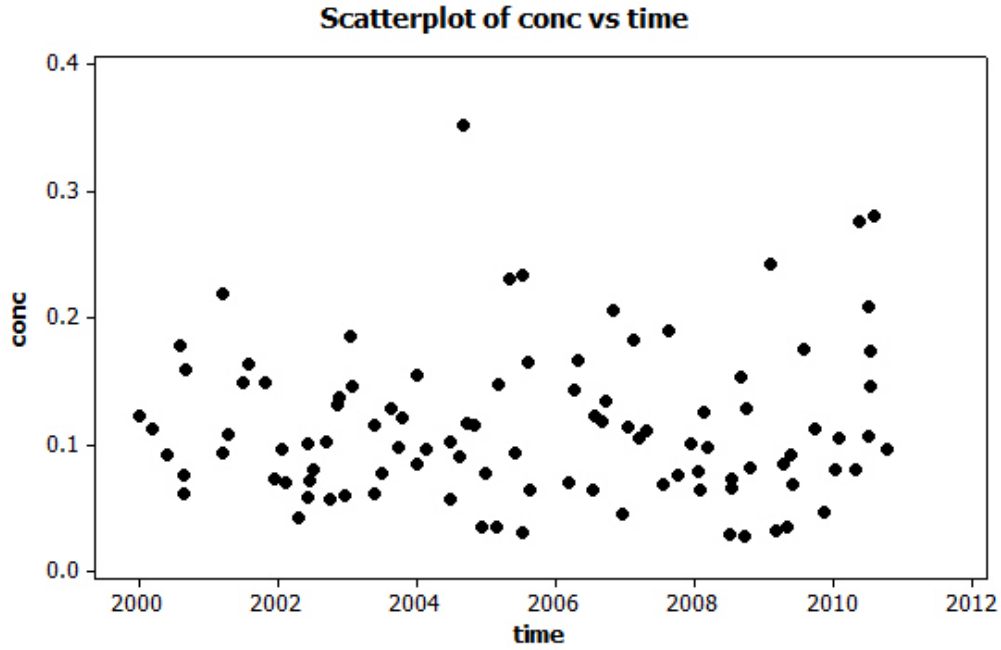
- Linear data. Should provide a good regression estimate.
- Censored 60% of data at two detection limits, 3 and 1 ppb
- Then substituted one-half the DL

Example 1: The Invasive Pattern



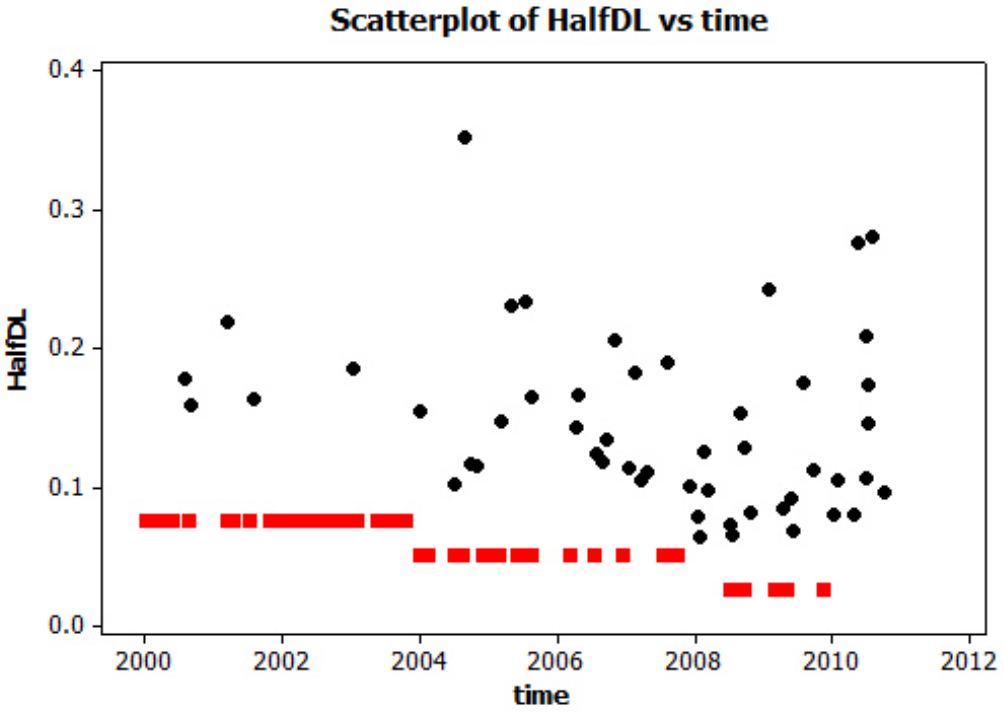
- The 60% censored data have a flat horizontal pattern of substituted values. This waters down the regression slope, making it too close to 0.
- Because many data have the same value, their standard deviation is too small, affecting the test of whether x is a significant predictor of y.

Example 2: Trend analysis



- Field data show no trend – true situation
- Will make some of the smallest values NDs, and substitute with 1/2DL. Detection limits decrease.

Example 2: After substitution, may 'find' a trend that's not there



- Above DL
- Below DL

Invasive pattern:

- DL decreases over time. After substitution with 1/2DL, the "data" decrease with time and the "trend" may become significant

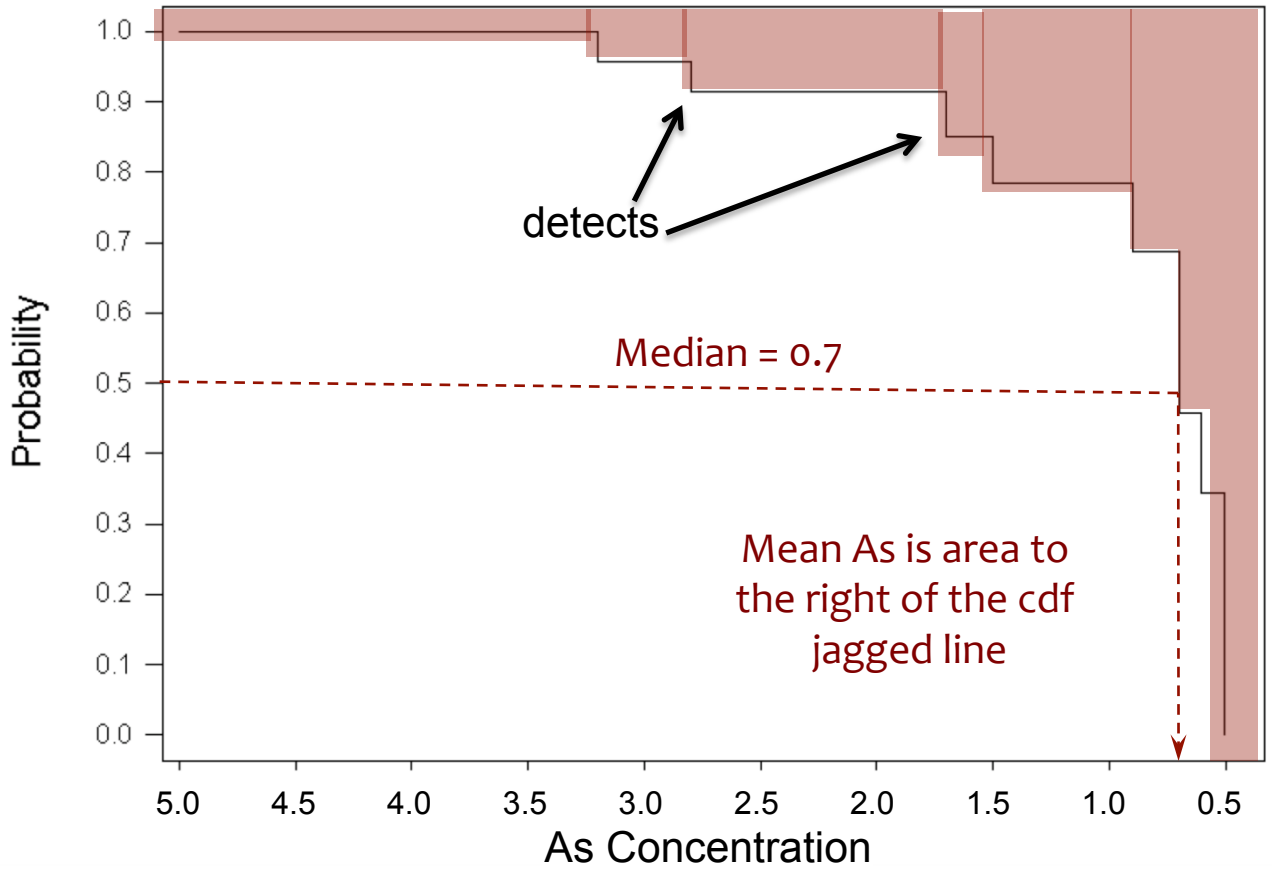
Example 3: Computing mean and std dev

Substitution will produce different results depending on what is substituted.

Value Subbed	Mean	StDev	Pct25	Median	Pct75
Zero	0.567	0.895	0.000	0.000	0.700
1/2 dl	1.002	0.699	0.500	0.950	1.000
dl	1.438	0.761	0.750	1.250	2.000

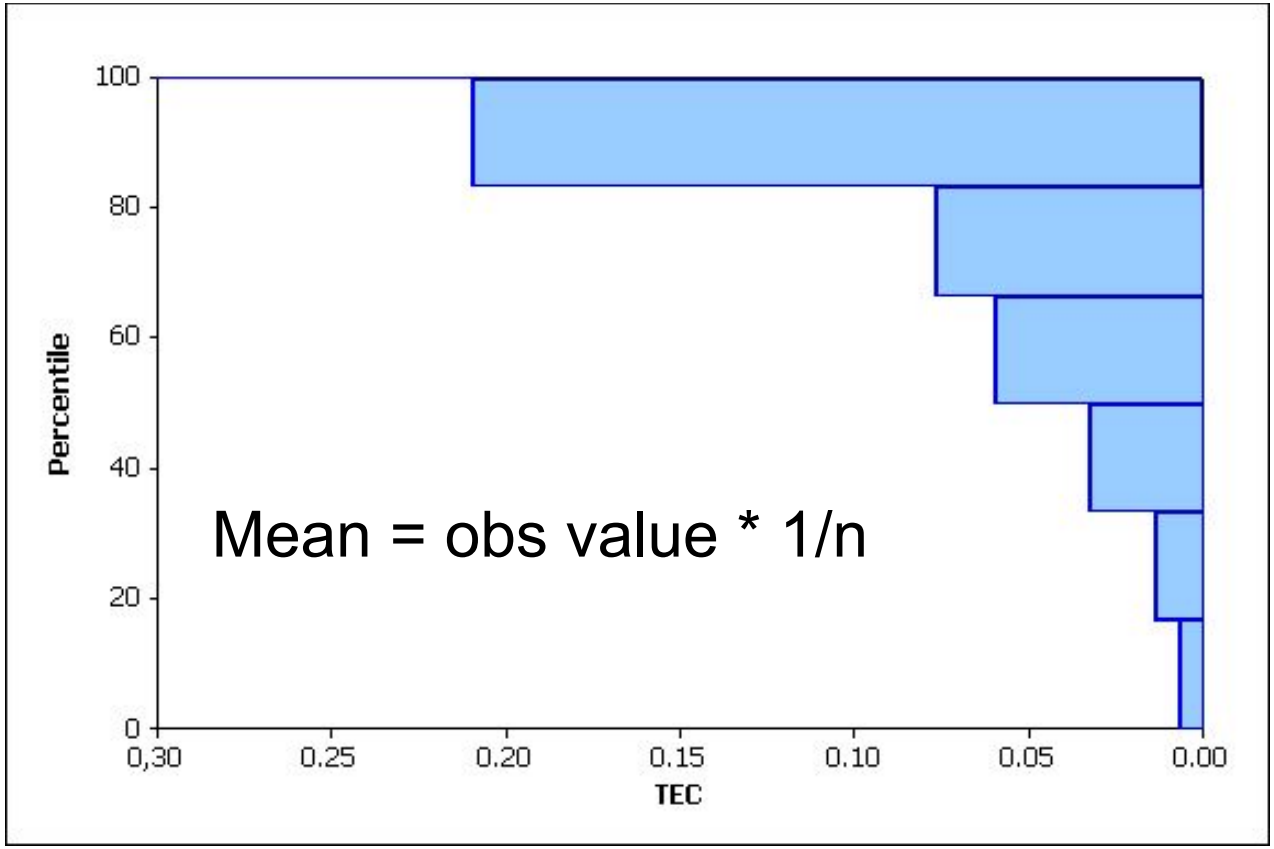
For the standard deviation (StDev), the range between zero and the DL does not necessarily include the true value. StDev is used in all parametric tests such as a t-test or regression. These tests may therefore be off the mark in a variety of ways if substitution is used.

Better method: Kaplan-Meier

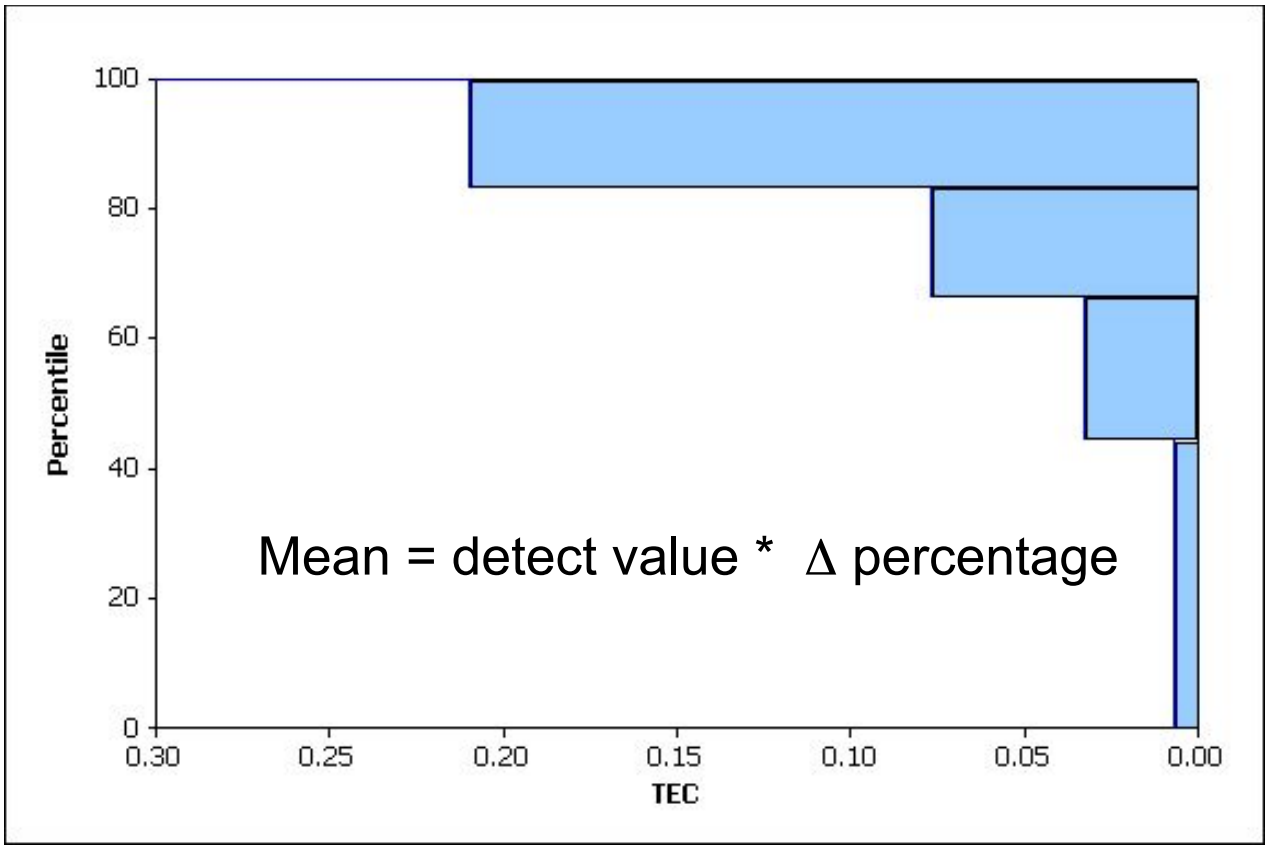


- Nonparametric, no substitution
- Percentiles for detects are adjusted for the number of nondetects.

Without nondetects, each obs has a weight of 1/n (thickness of each bar) when computing the mean



Kaplan-Meier: detects are unequally weighted by # of points (detects and NDs) above and below that value



Summary: Statistics for Censored Data

- Substitution may give wrong results!
- Kaplan-Meier and other survival analysis methods provide the full solution for one or several detection limits
- For one DL familiar and simpler nonparametric tests (rank-sum, Kruskal-Wallis test...) work well.
- Neither insert invasive patterns.
- Standard procedures that can easily be included in regulations

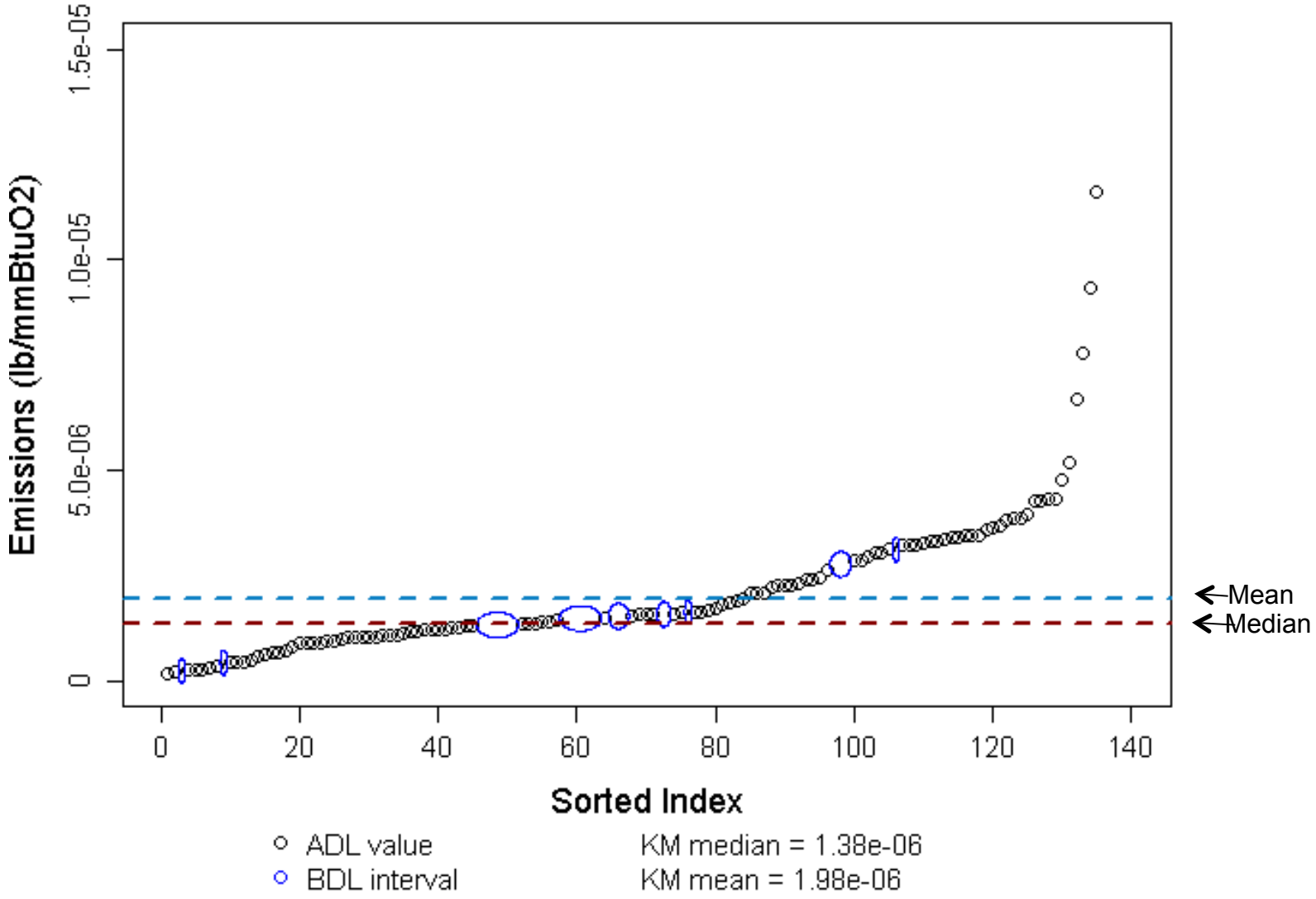
Application to EGU ICR Emissions Data

- Analyzed a subset of 2010 HAPs ICR Data
  - Stack Emissions from Oil-fired Units: Se, Be, HF
- ICR testing protocol
  - Three runs per test
  - Report all concentrations Above Detection Level (ADL)
  - Report non-detects as
    - BDL if all data points Below Detection Levels
    - DLL – in multi-fraction analyses, if one fraction is above, another below detection level (Detection Level Limited)
- Computed Kaplan-Meier means and medians using 'R', a public domain statistical software package
- Compared with arithmetic means and standard deviations using simple substitution for non-detects

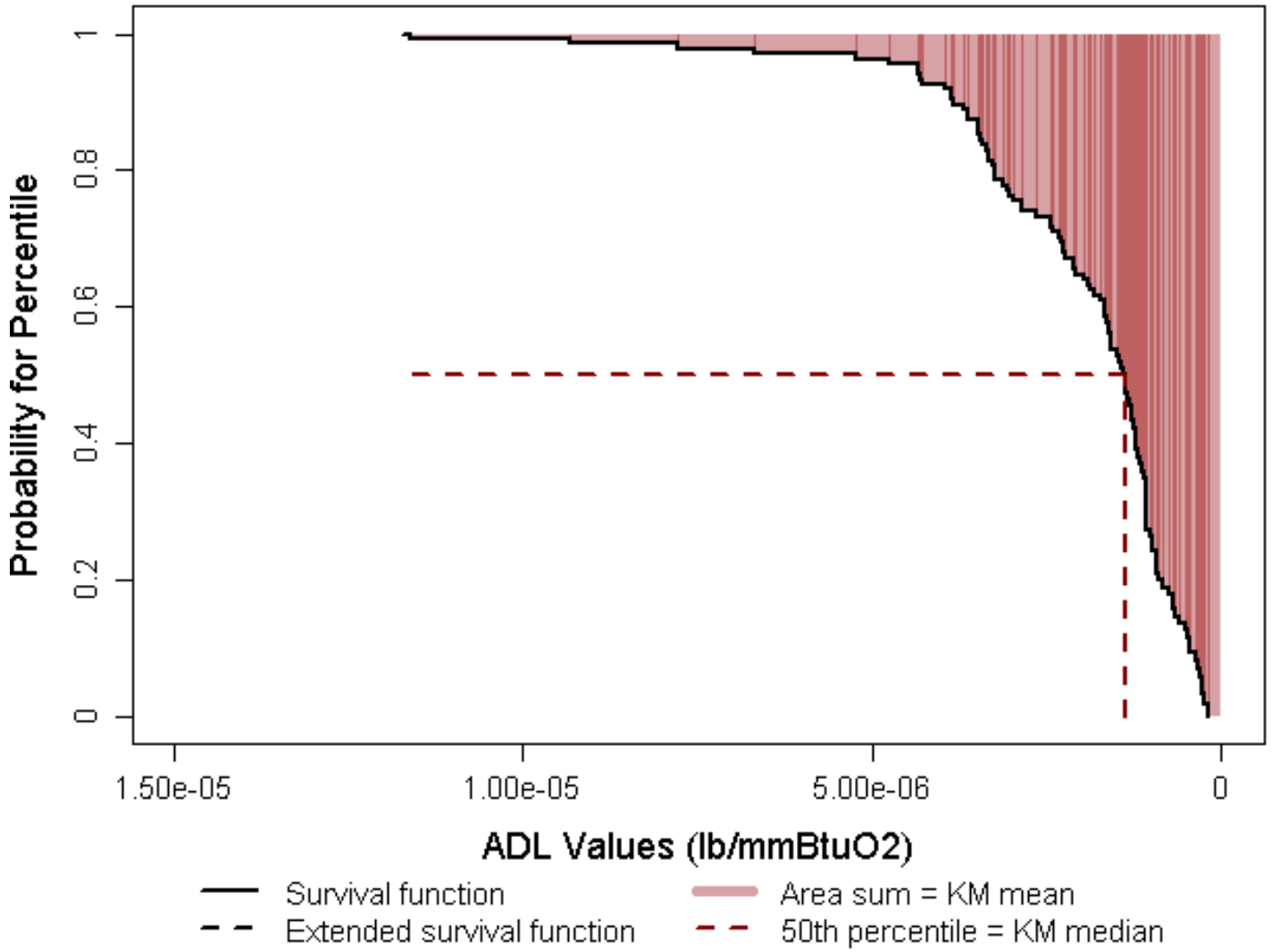
2010 HAPs ICR Stack Emissions Data
Oil-fired Units – Selenium

Substitution	Emissions (lb/trillion Btu)					
	#6 Oil Uncontrolled (135 total; 24 censored)		#6 Oil w/ESP (26 total; 9 censored)		#2 Oil (18 total; 6 censored)	
	Arithmetic Mean	Std. Deviation	Arithmetic Mean	Std. Deviation	Arithmetic Mean	Std. Deviation
Zero	1.83	1.83	1.68	1.45	0.28	0.39
½ DL	1.97	1.71	1.90	1.22	0.50	0.36
DL	2.11	1.66	2.11	1.07	0.73	0.57
None **(Kaplan-Meier**)	Mean =1.98 Median = 1.38	1.72	Mean = 1.90 Median = 2.06	1.24	Mean = 0.42 Median= 0.18	0.42

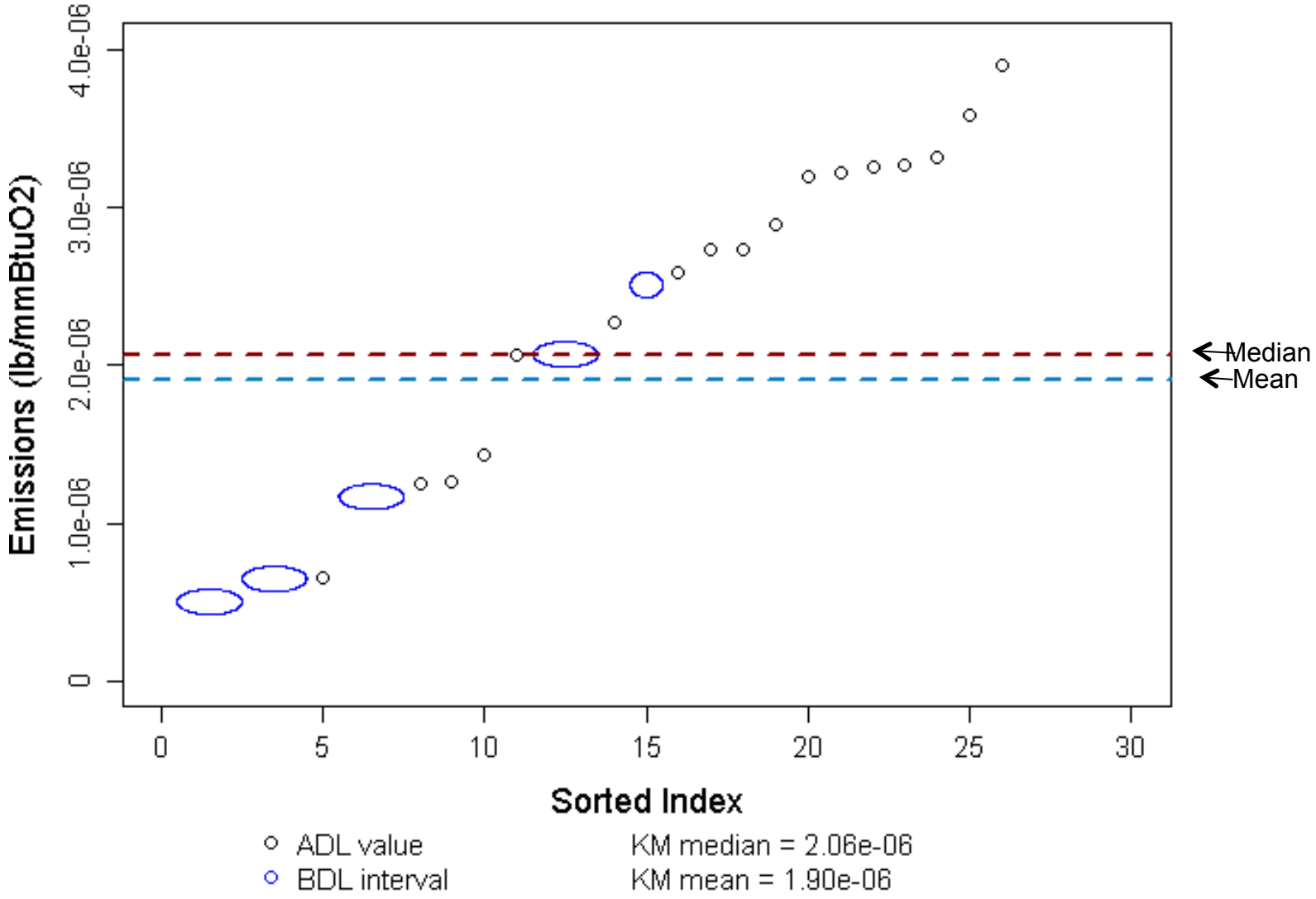
Selenium – Uncontrolled #6 Oil-fired Units Emissions Data



Selenium – Uncontrolled #6 Oil-fired Units K-M Survival Function Plot

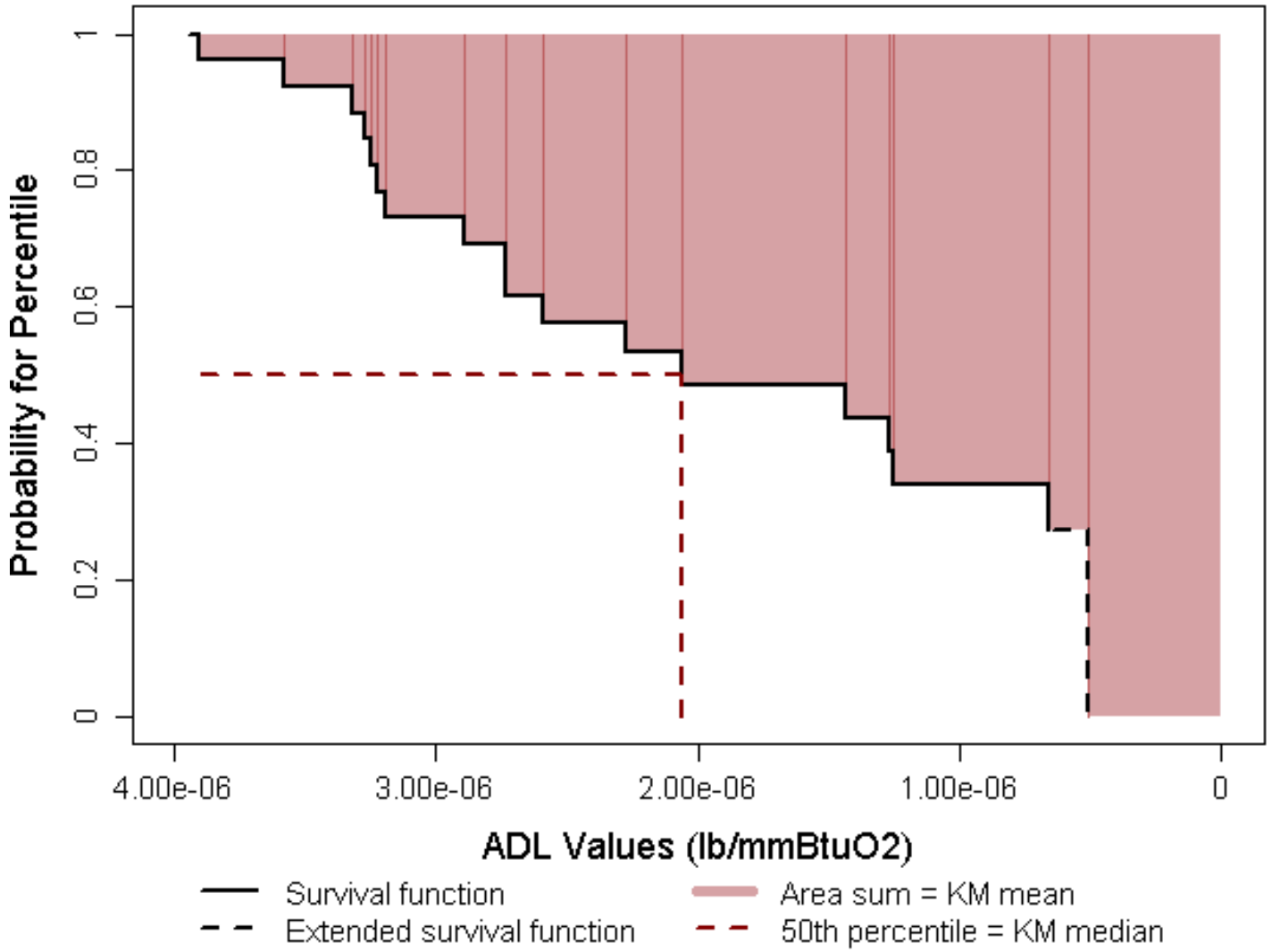


Selenium - #6 Oil-fired Units with ESP Emissions Data

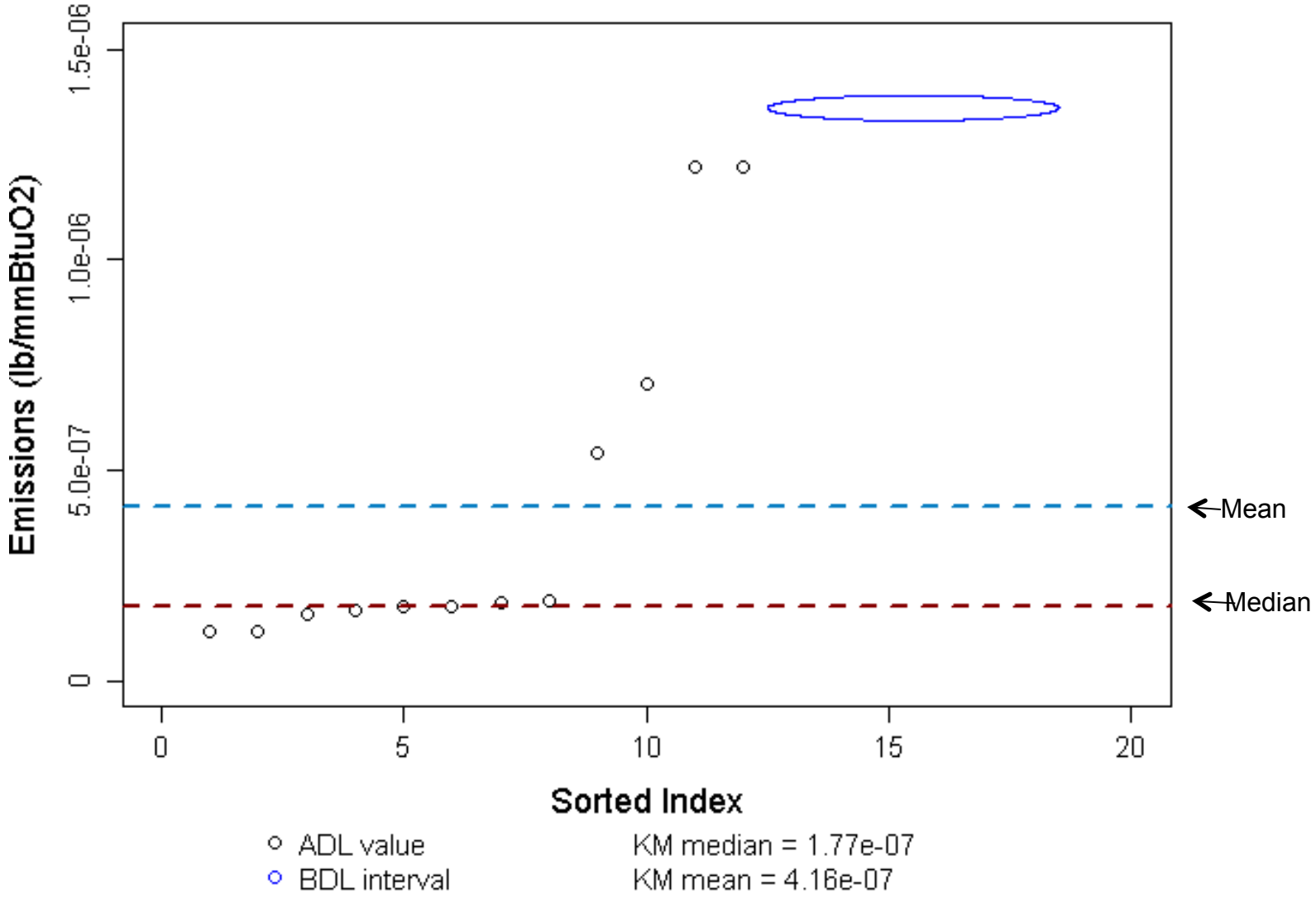


EPRI ELECTRIC POWER RESEARCH INSTITUTE

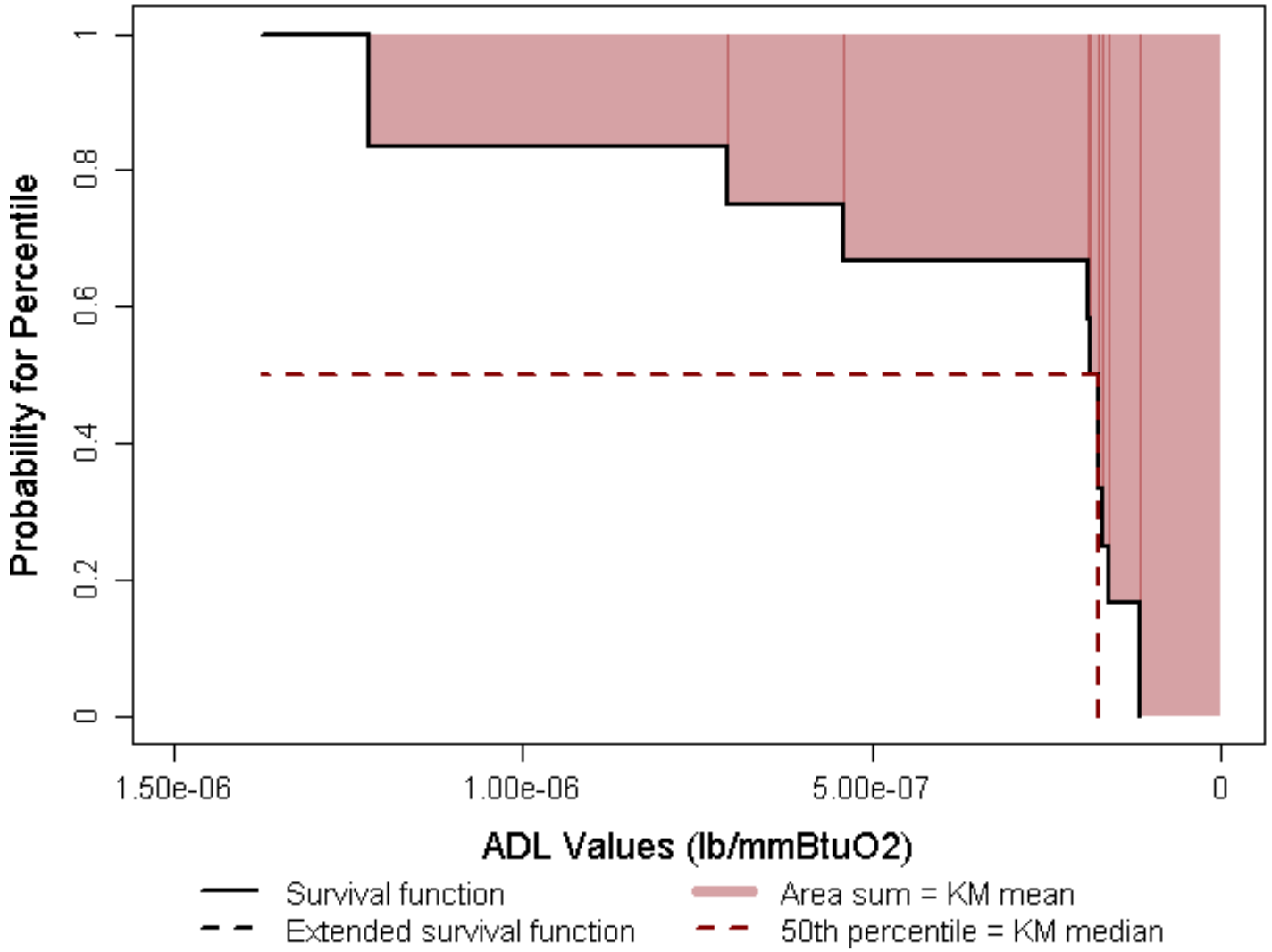
Selenium - #6 Oil-fired Units with ESP
K-M Survival Function Plot



Selenium - #2 Oil-fired Units (Uncontrolled) Emissions Data



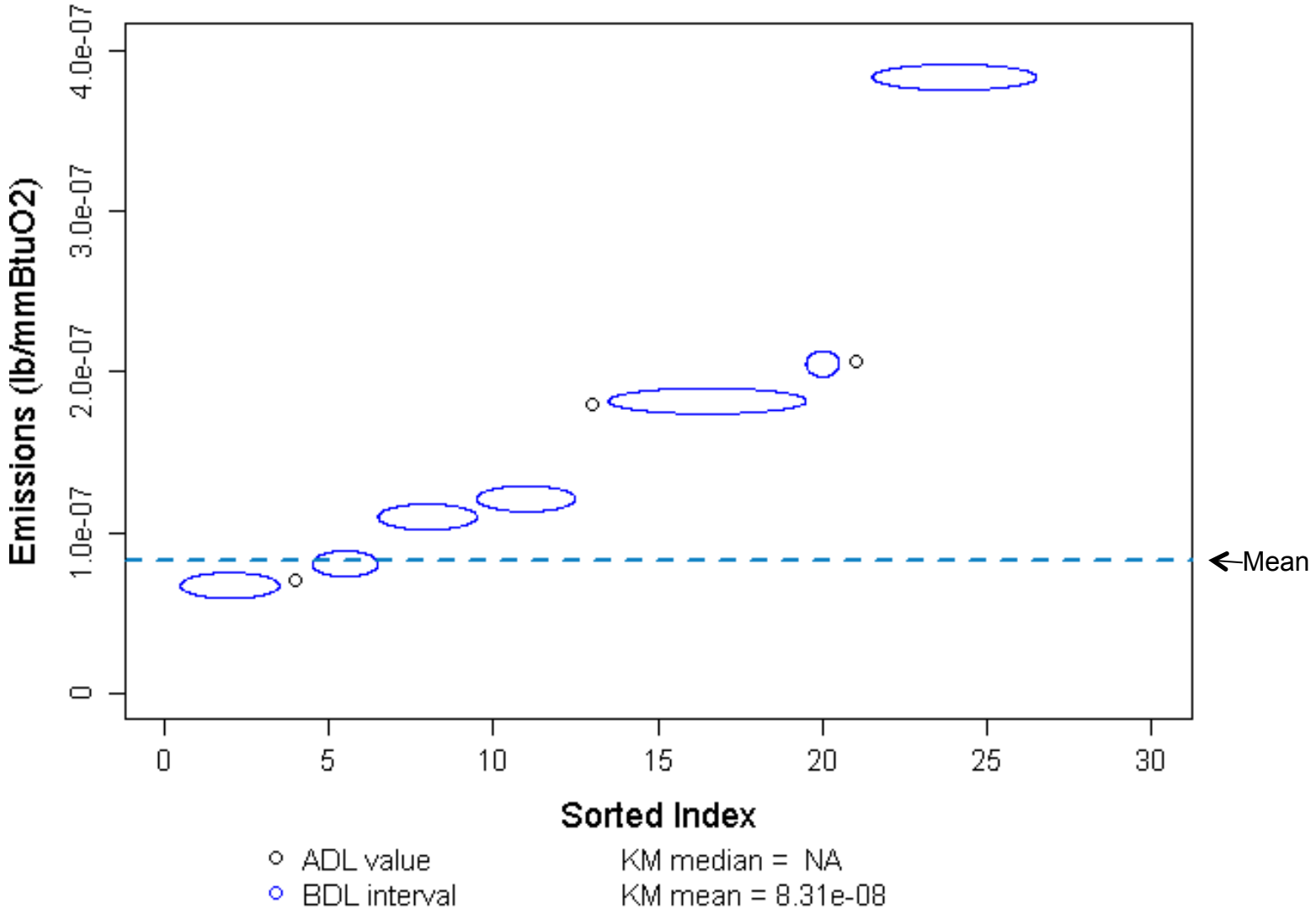
Selenium - #2 Oil-fired Units (Uncontrolled) K-M Survival Function Plot



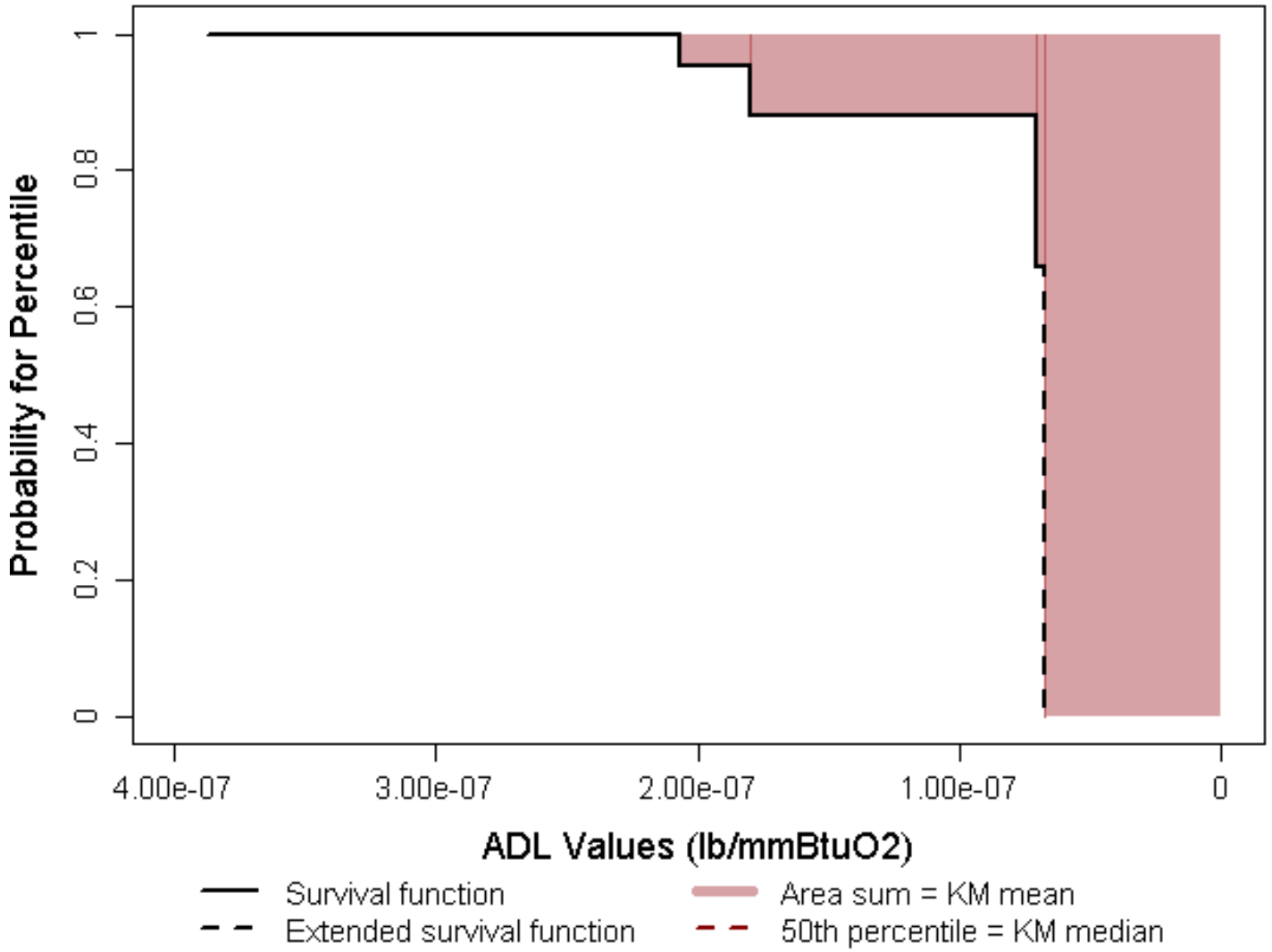
2010 HAPs ICR Stack Emissions Data
Oil-fired Units – Beryllium

Substitution	Emissions (lb/trillion Btu)			
	#6 Oil Uncontrolled (135 total; 28 censored)		#6 Oil w/ESP (26 total; 23 censored)	
	Arithmetic Mean	Std. Deviation	Arithmetic Mean	Std. Deviation
Zero	0.233	0.228	0.018	0.054
½ DL	0.255	0.210	0.100	0.061
DL	0.276	0.201	0.182	0.110
None (**Kaplan-Meier**)	Mean = 0.251 Median = 0.165	0.216	Mean = 0.0831 Median = NA	0.050

Beryllium – #6 Oil-fired Units with ESP Emissions Data



Beryllium – #6 Oil-fired Units with ESP
K-M Survival Analysis Plot



2010 HAPs ICR Stack Emissions Data
Oil-fired Units – HF

Substitution	Emissions (lb/trillion Btu)			
	#6 Oil Uncontrolled (73 total; 46 censored)		#6 Oil w/ESP (17 total; 1 censored)	
	Arithmetic Mean	Std. Deviation	Arithmetic Mean	Std. Deviation
Zero	172	531	135	44
½ DL	238	514	135	42
DL	304	505	136	41
None (**Kaplan-Meier**)	Mean = 217 Median = 61	527	Mean = 136* Median = 136* * K-M not suitable	41

Conclusions

- Substitution methods are arbitrary; can lead to erroneous conclusions
- Methods are available for computing descriptive statistics, plots, hypothesis tests, and regression for censored data, all without substitution
- Binary and familiar nonparametric methods can be used for data with 1 DL
- Demonstrated application of non-parametric Kaplan-Meier method to selected emissions data from power plants
- Recommend use of more sound statistical methods rather than simple substitution when dealing with environmental data containing non-detects