Statistical Analysis of Environmental Data

Presented by The Palladino Company, Inc.

Presented for the U.S. Environmental Protection Agency, Region 9

Session Code: W-F1

Date: March 26, 2025

Instructor

Carl Palladino
Health Physicist
415-336-1556
carl@palladinocompany.com

Course Sponsor

Robert Wise
Federal On-Scene Coordinator
562-889-2572
wise.robert@epa.gov

Course Agenda

Data Distribution

Normal, lognormal, gamma, and non-parametric distributions in environmental data

Outliers

Identification methods and impact analysis

UCLs and BTVs

Upper confidence limits and background threshold values

Hypothesis Testing

Statistical decision-making

Decision Errors

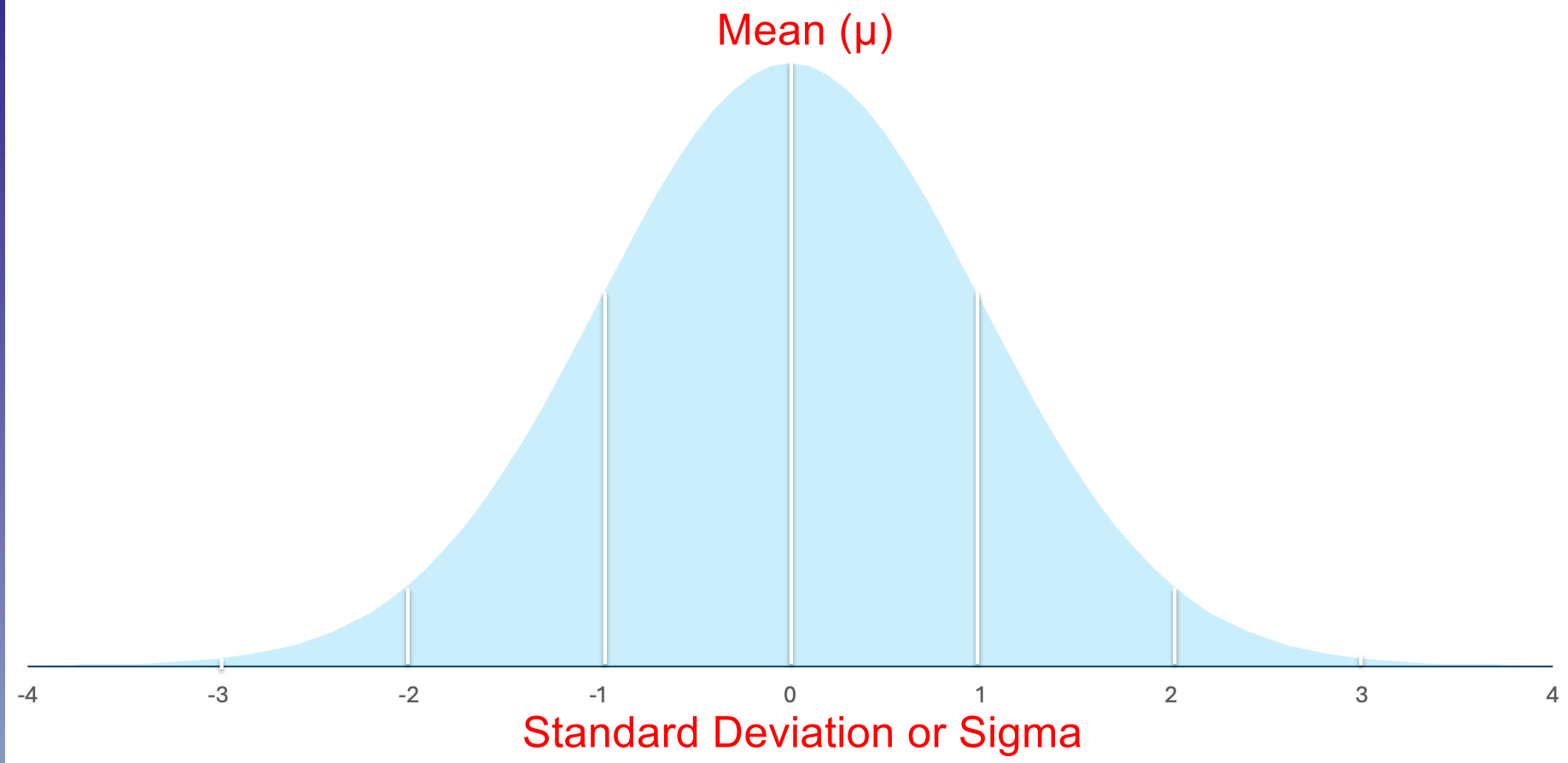
Reducing decision errors

ProUCL Software

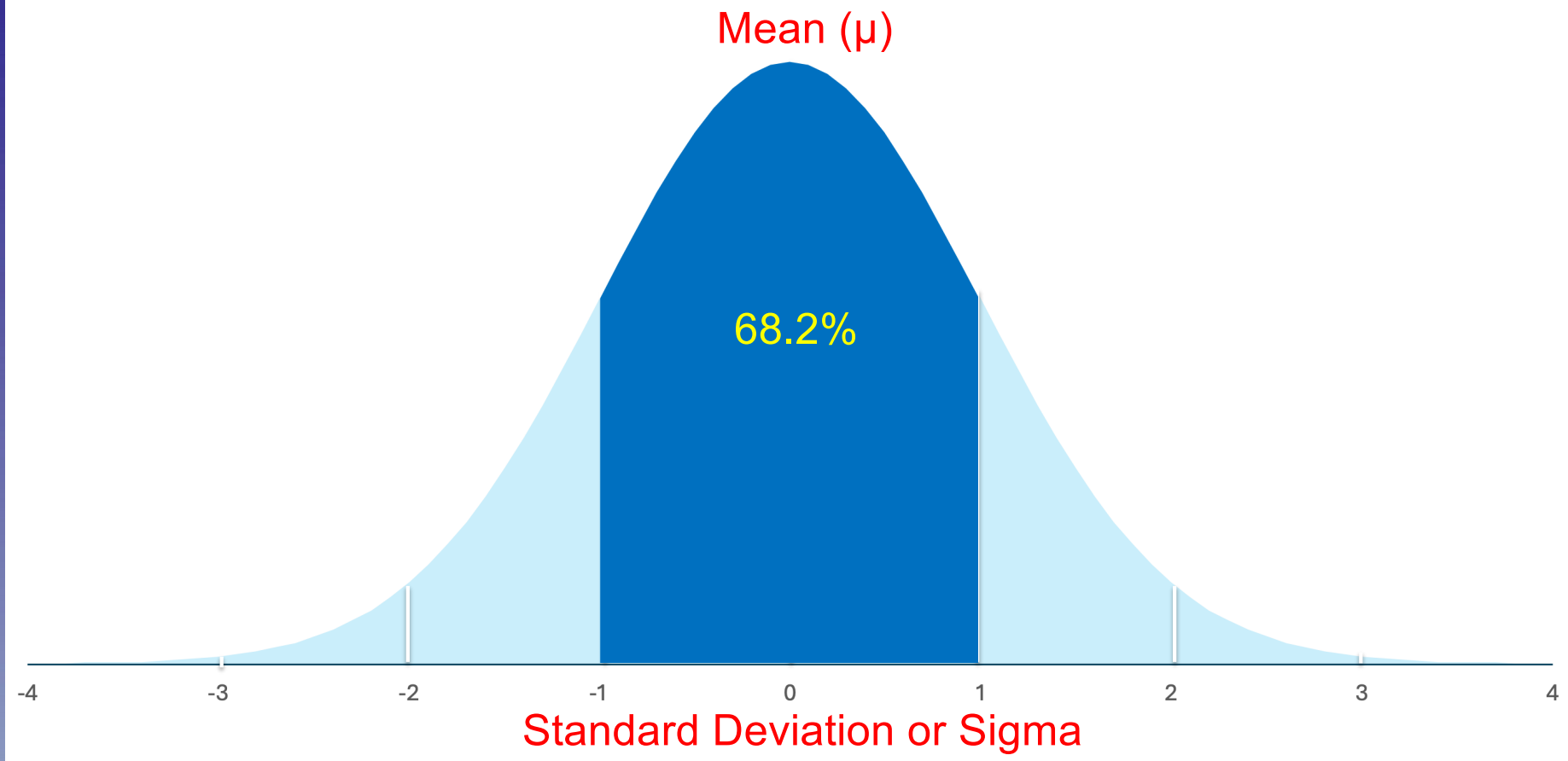
Statistical software for environmental data

Data Distributions

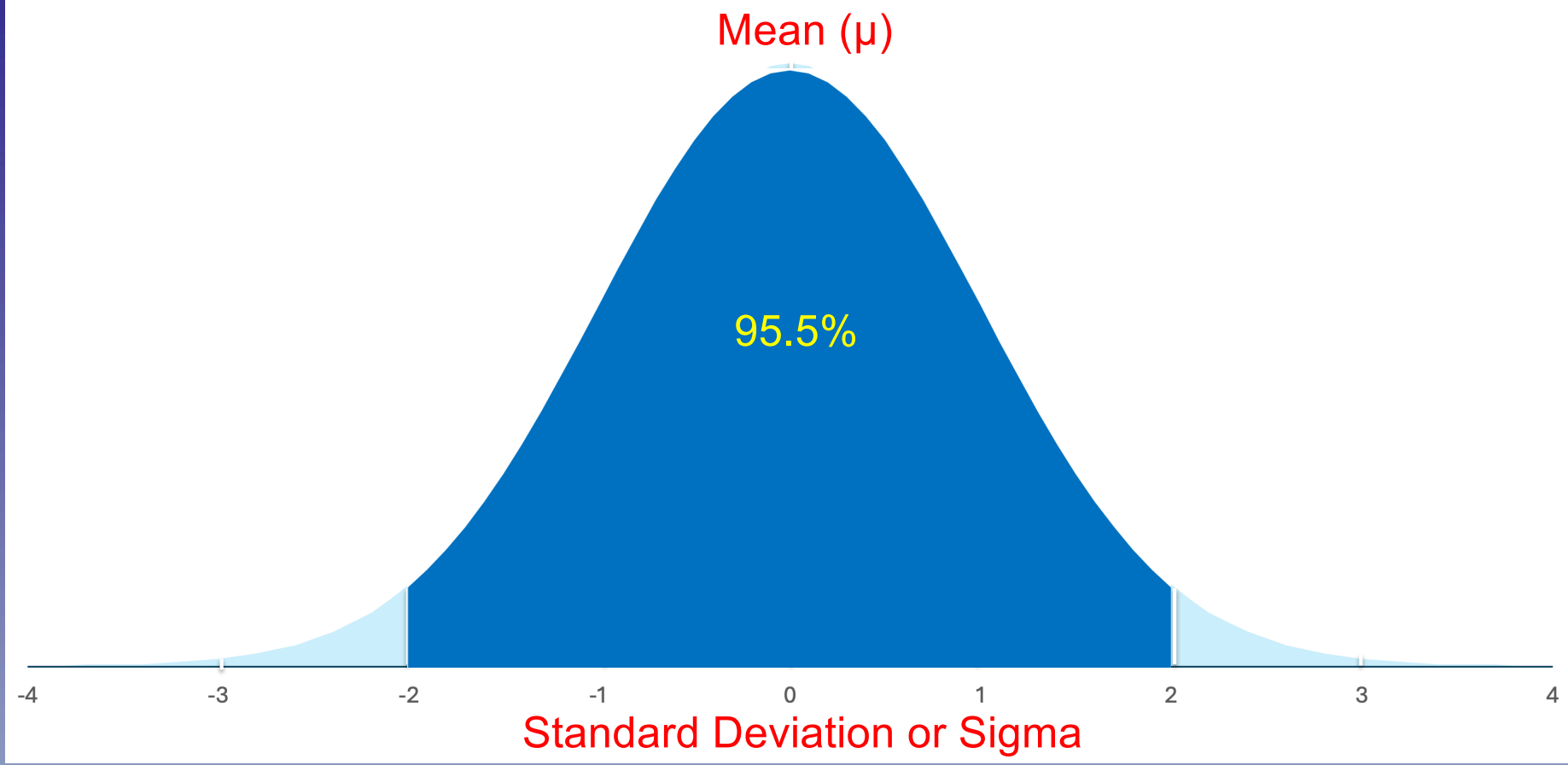
Normal Distribution



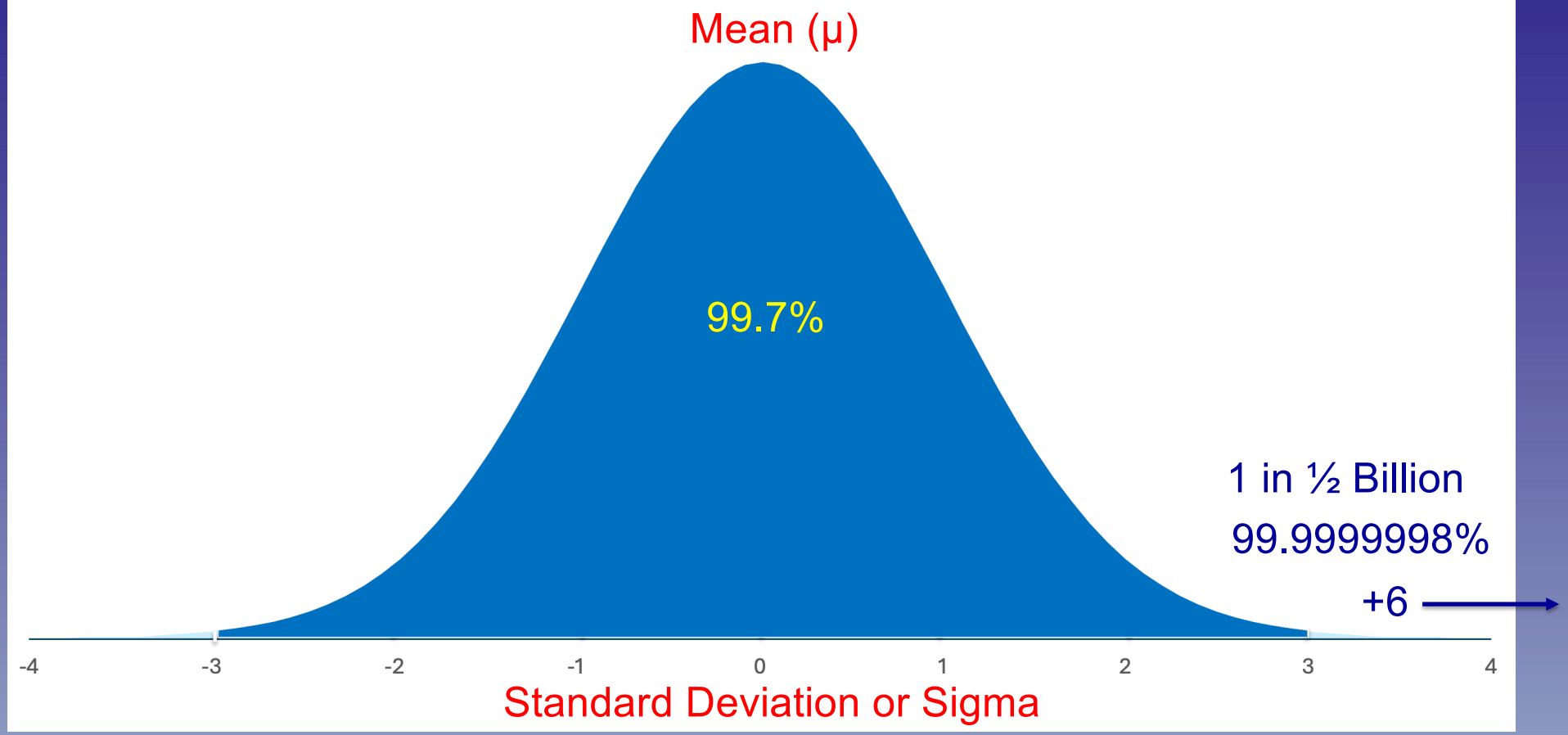
Normal Distribution



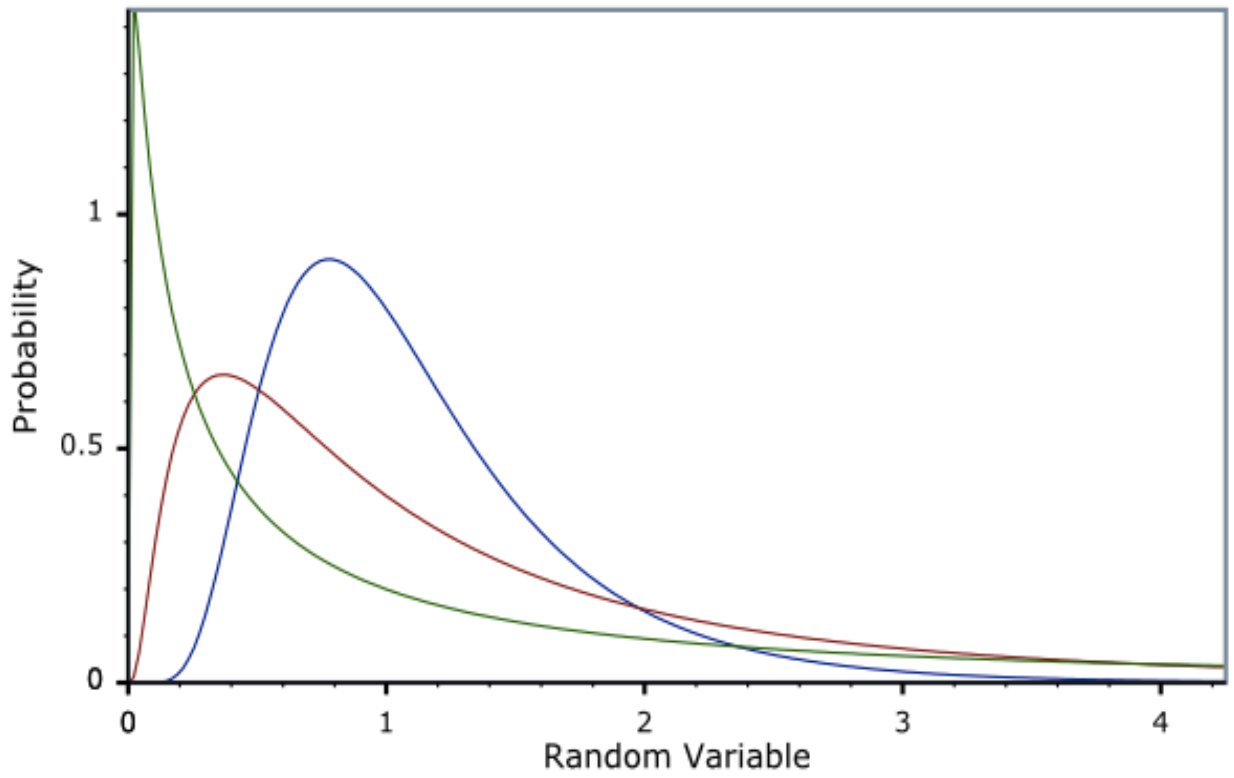
Normal Distribution



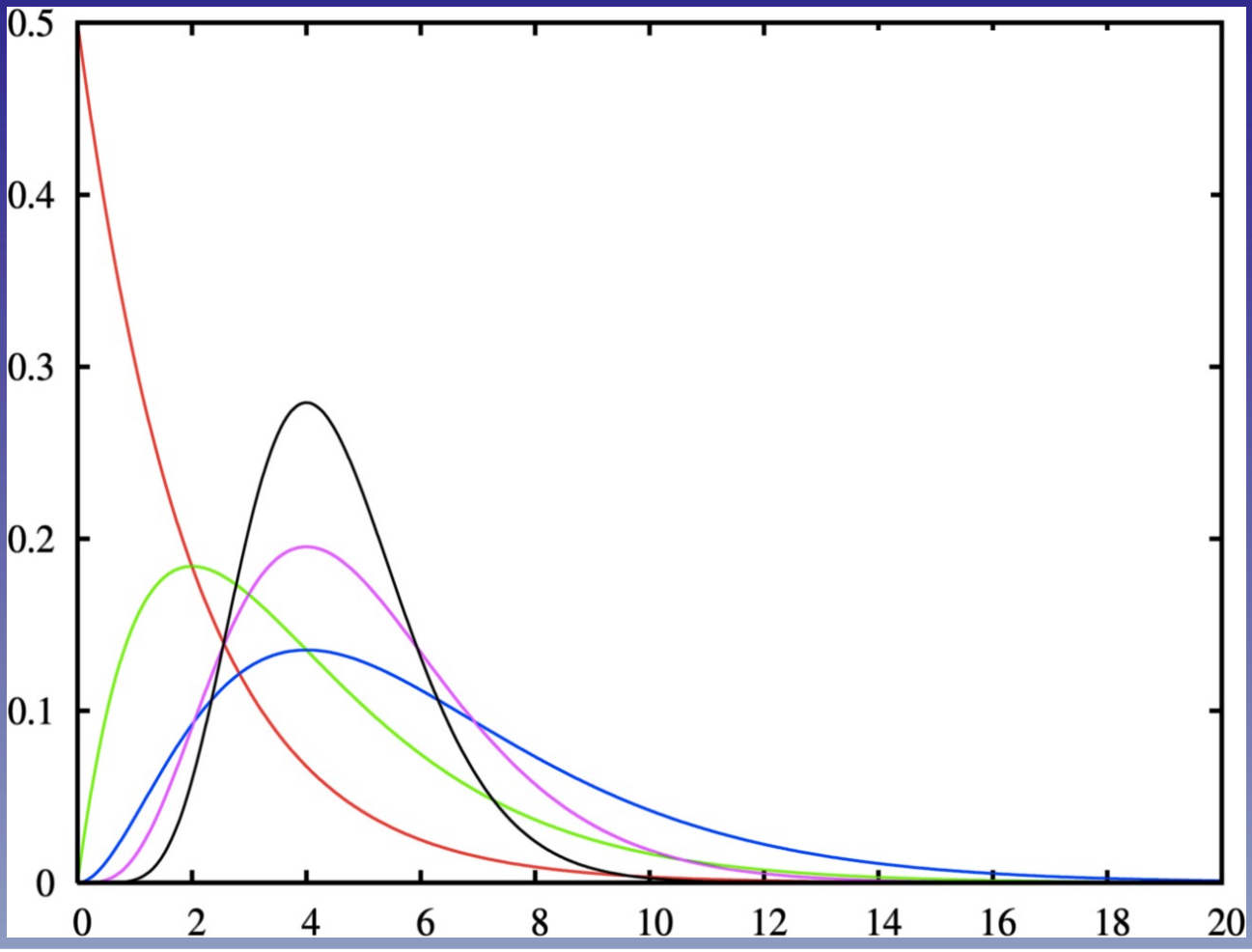
Normal Distribution



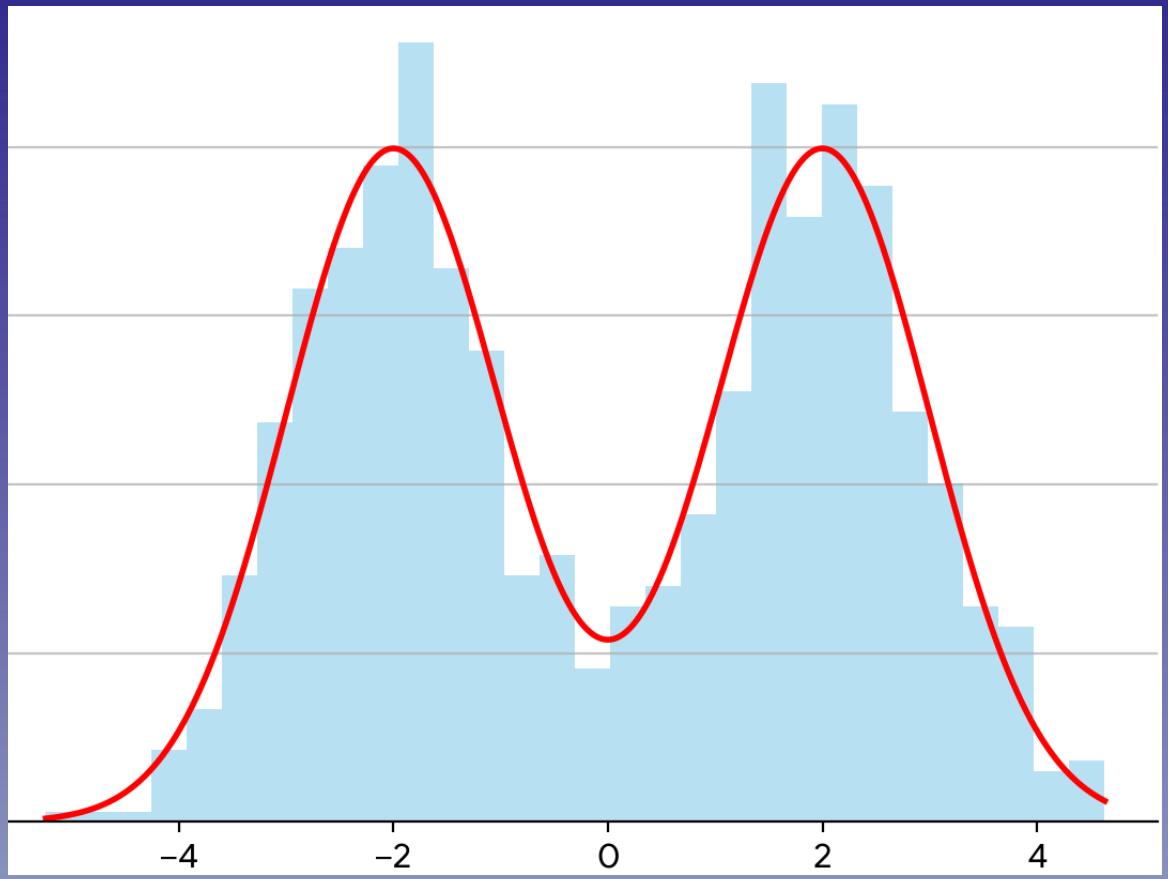
Lognormal Distributions



Gamma Distributions

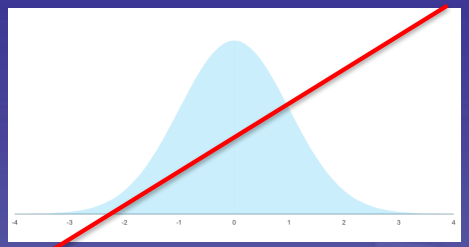


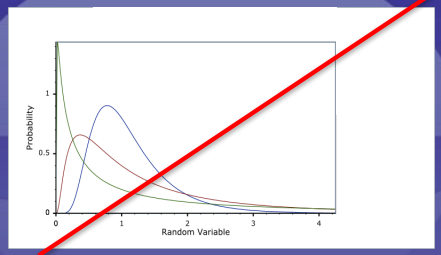
Watch Out for Bimodal Distributions

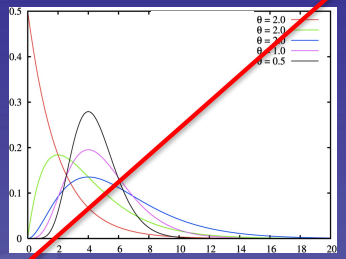


Nonparametric Distribution

- Data does not fit a distribution







- Nonparametric statistics do not assume predefined distribution parameters

- Downside to nonparametric statistics is reduced power

Power: probability a statistical test correctly rejects the null hypothesis

Example Data Set

Ra-226 Background Data Set

1)	0.740	11)	7.39
2)	1.24	12)	8.40
3)	1.75	13)	8.40
4)	2.25	14)	9.43
5)	2.28	15)	9.47
6)	3.33	16)	10.51
7)	3.36	17)	10.52
8)	5.35	18)	13.56
9)	7.36	19)	22.11
10)	7.38	20)	35.87

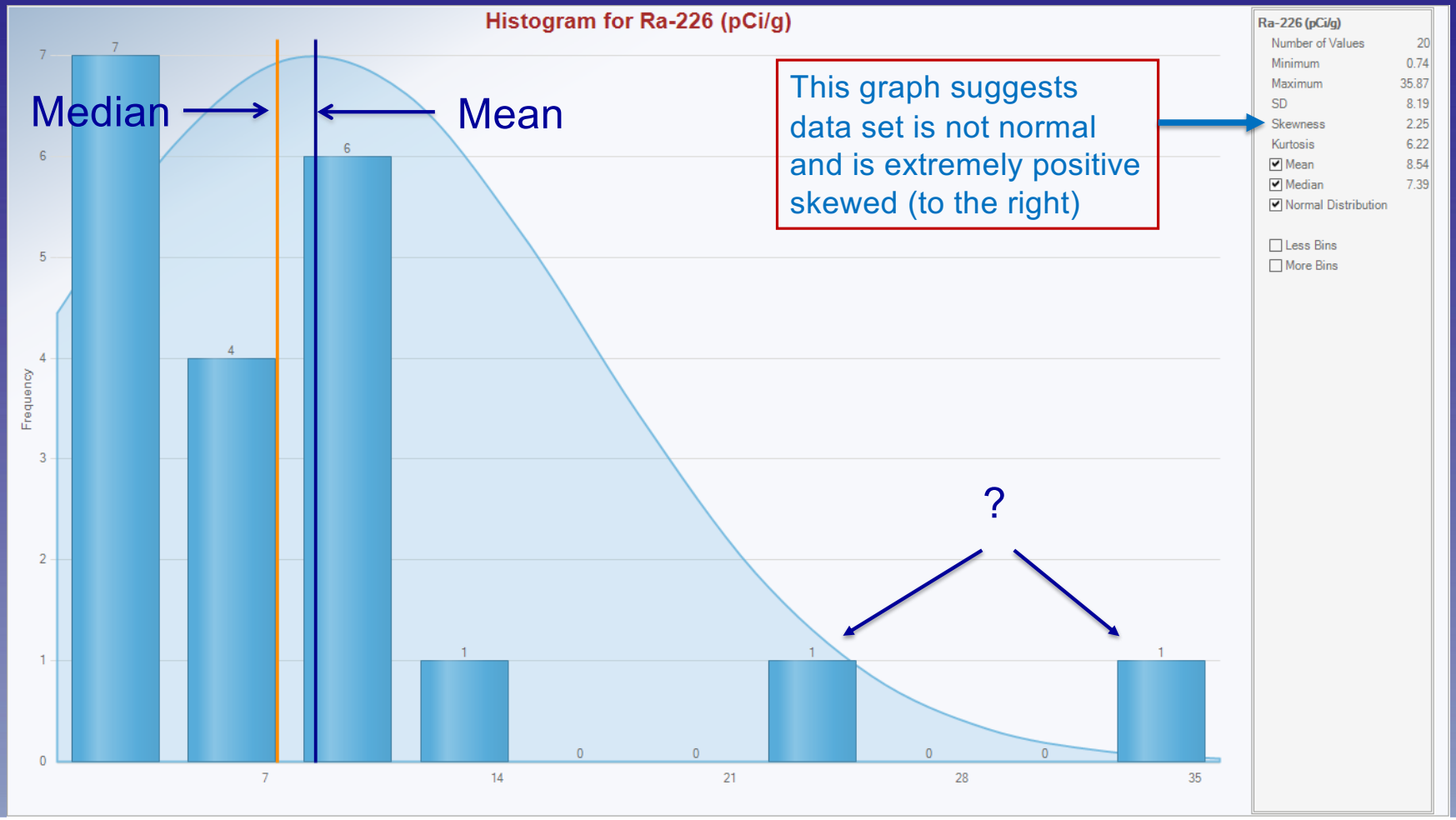
n = 20

Mean = 8.53

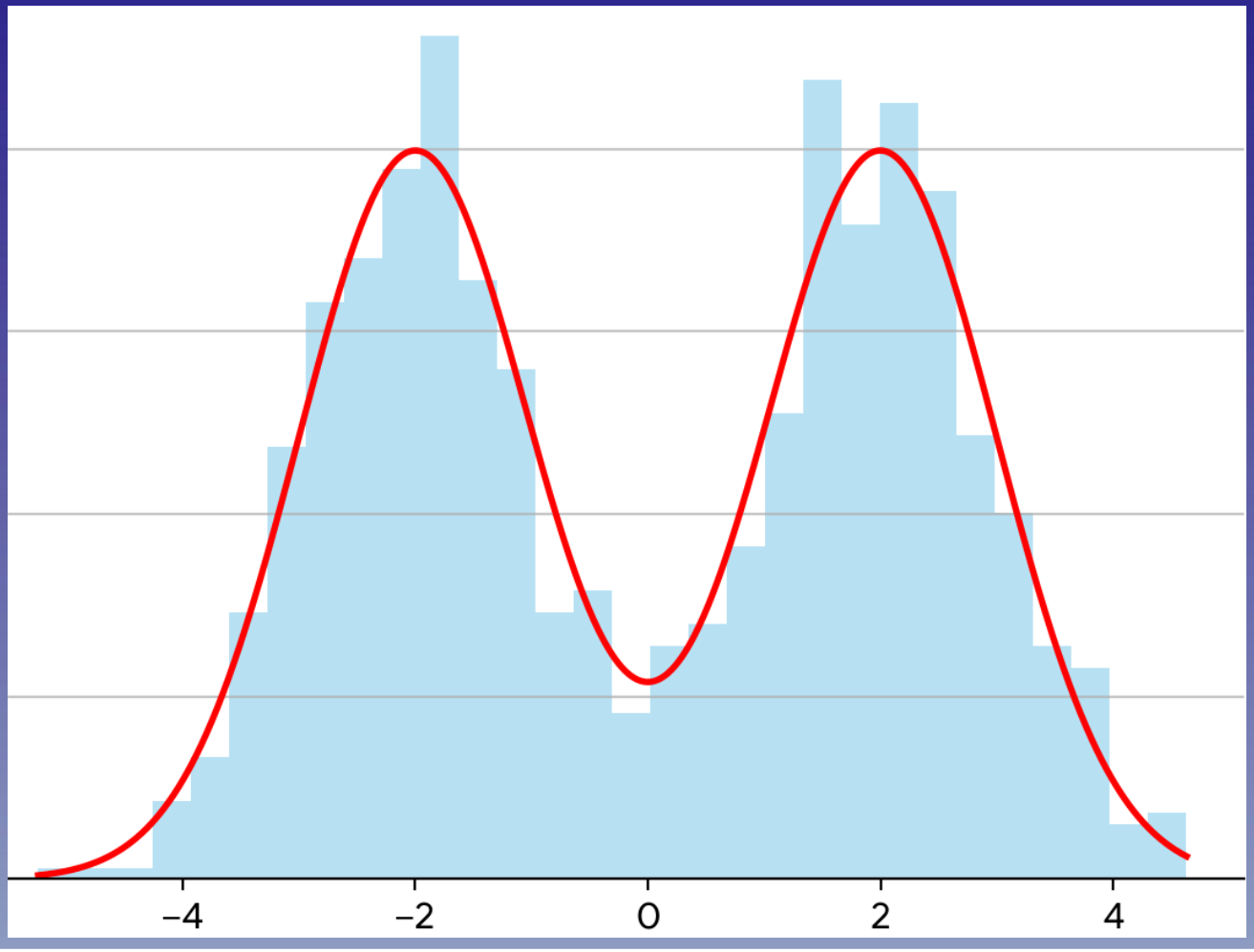
Median = 7.38

Standard Deviation = 8.18

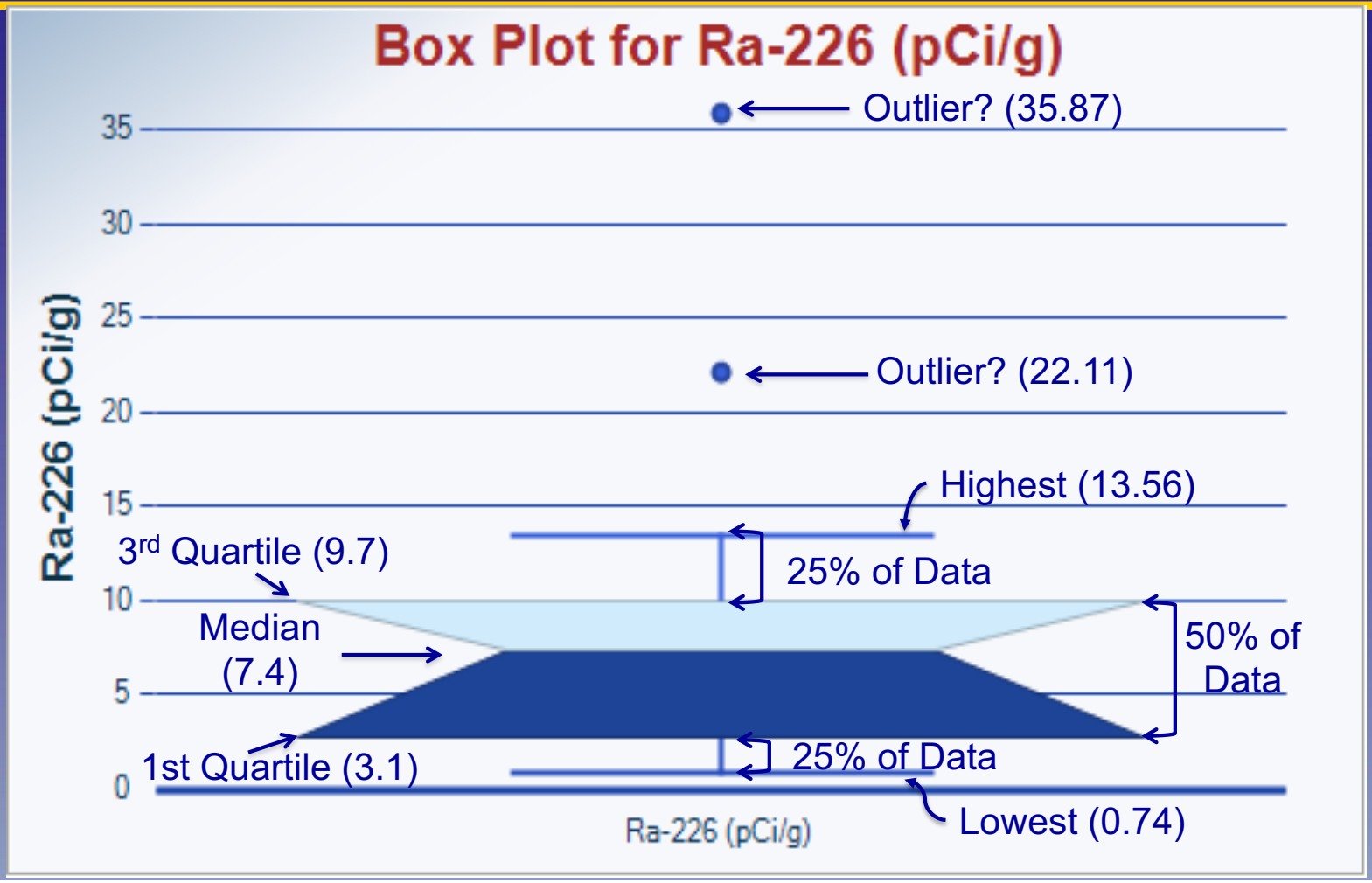
Histogram


Histogram for Ra-226 (pCi/g)
Median
Mean
This graph suggests data set is not normal and is extremely positive skewed (to the right)
?
Frequency
Ra-226 (pCi/g)
Number of Values 20
Minimum 0.74
Maximum 35.87
SD 8.19
Skewness 2.25
Kurtosis 6.22
Mean 8.54
Median 7.39
Normal Distribution
Less Bins
More Bins

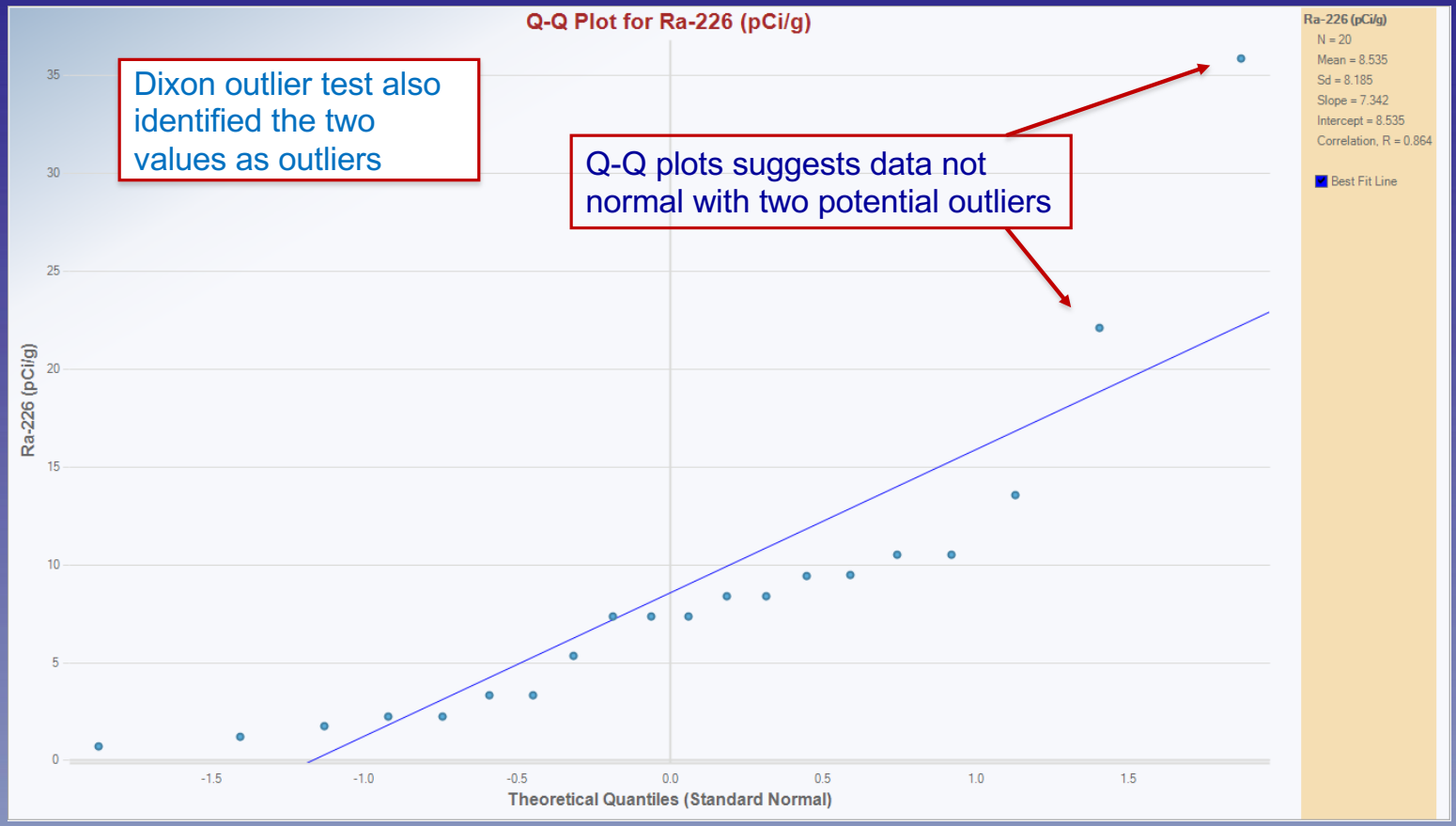
Watch Out for Bimodal Distributions



Box Plot


Box Plot for Ra-226 (pCi/g)
Outlier? (35.87)
Outlier? (22.11)
Highest (13.56)
3rd Quartile (9.7)
25% of Data
Median (7.4)
50% of Data
1st Quartile (3.1)
25% of Data
Lowest (0.74)
Ra-226 (pCi/g)
Ra-226 (pCi/g)
35
30
25
20
15
10
5
0

Quantile-Quantile (Q-Q) Plot

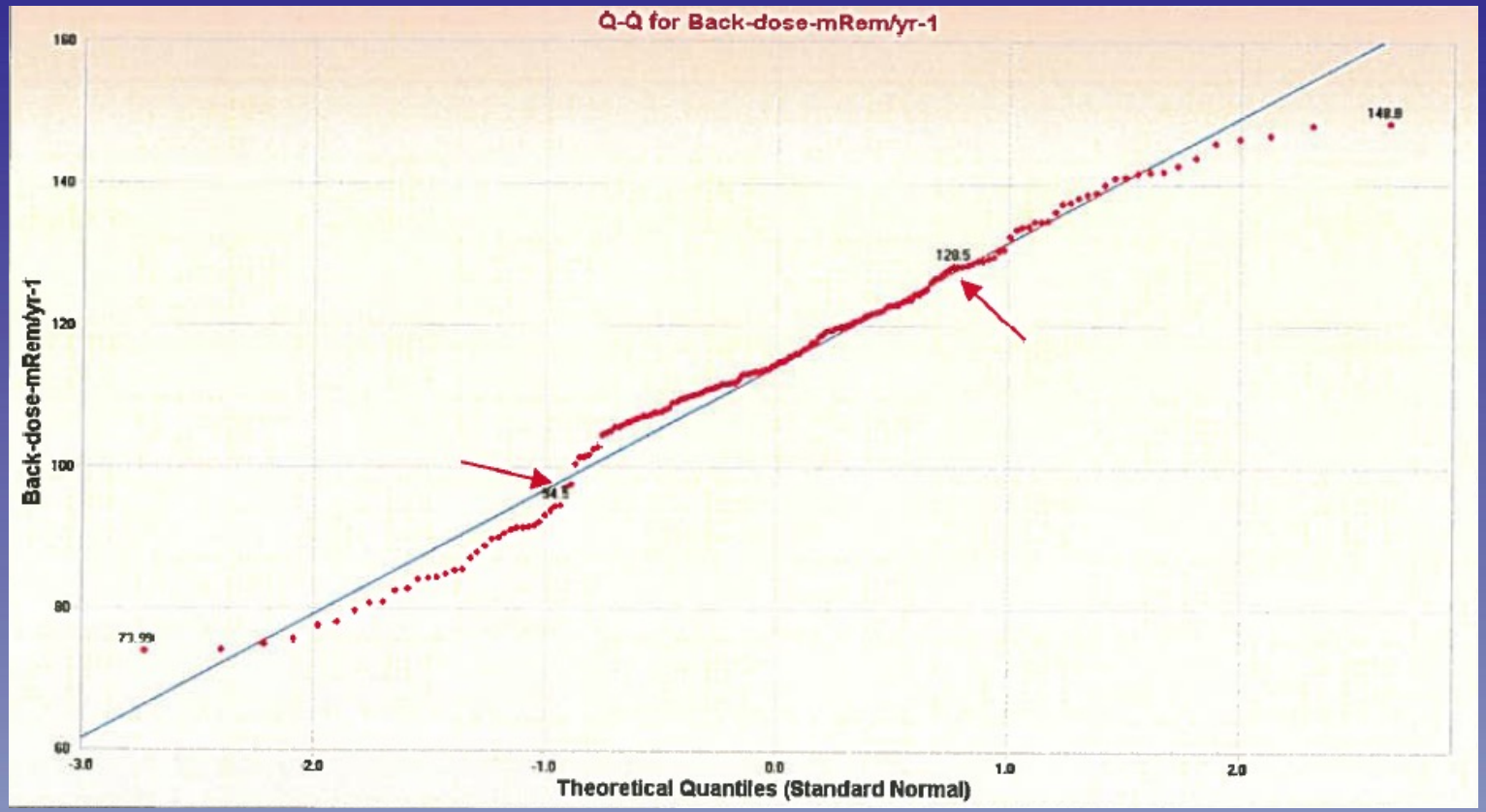


Goodness of Fit (GOF)

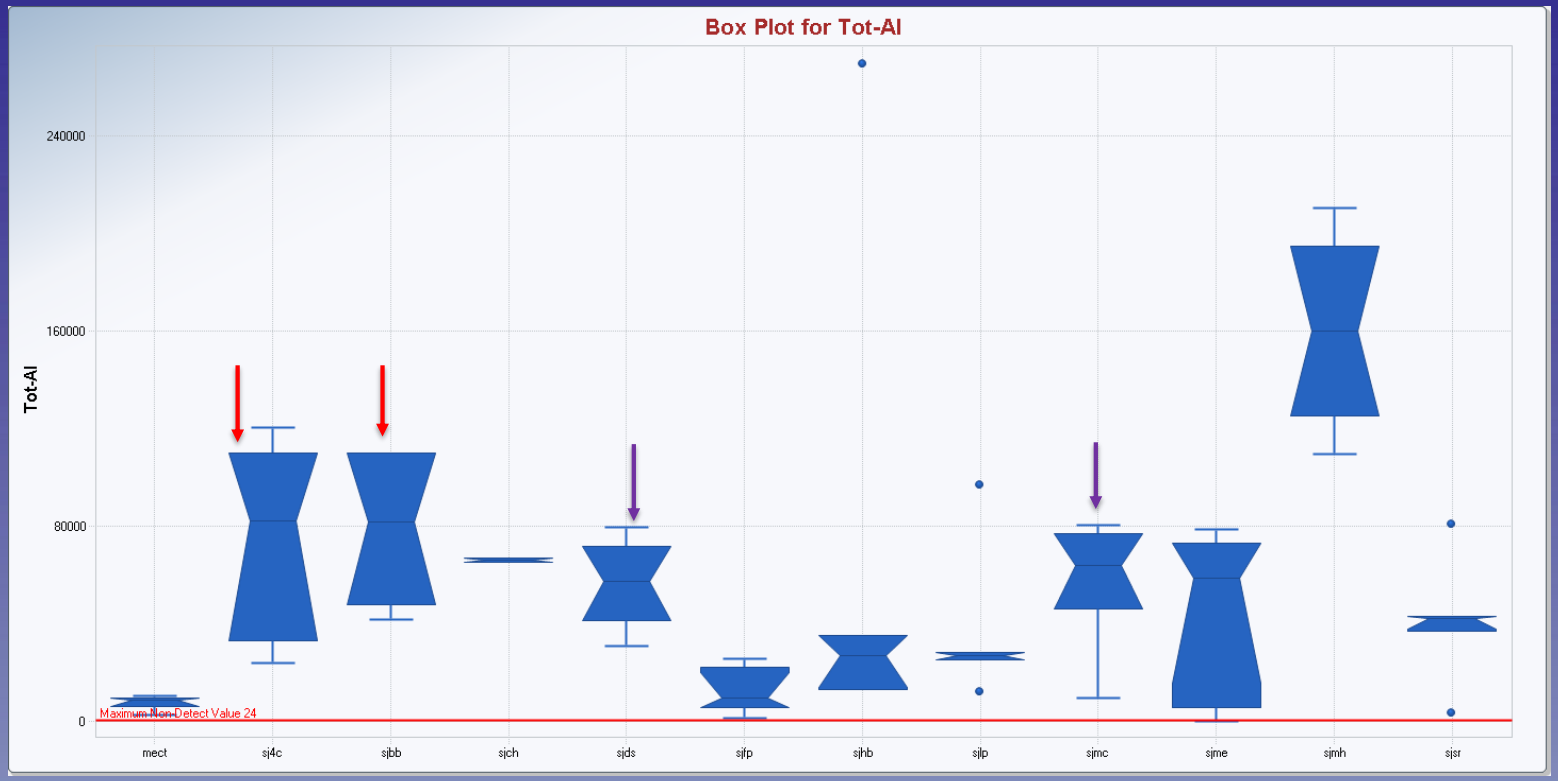
- GOF tests the probability that distribution fits a model; examples:
  - Normal: Shapiro Wilk and Lilliefors
  - Lognormal: Shapiro Wilk and Lilliefors
  - Gamma: Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S)

- Ra-226 data set indicates a 95% probability:
  - Not Normal
  - ~~Approximately Lognormal~~
  - Gamma distributed

Multiple Data Populations



Multiple Data Populations



Outliers

Remove Outliers

Ra-226 Background Data Set

1)	0.740	11)	7.39
2)	1.24	12)	8.40
3)	1.75	13)	8.40
4)	2.25	14)	9.43
5)	2.28	15)	9.47
6)	3.33	16)	10.51
7)	3.36	17)	10.52
8)	5.35	18)	13.56
9)	7.36	~~19)~~	~~22.11~~
10)	7.38	~~20)~~	~~35.87~~

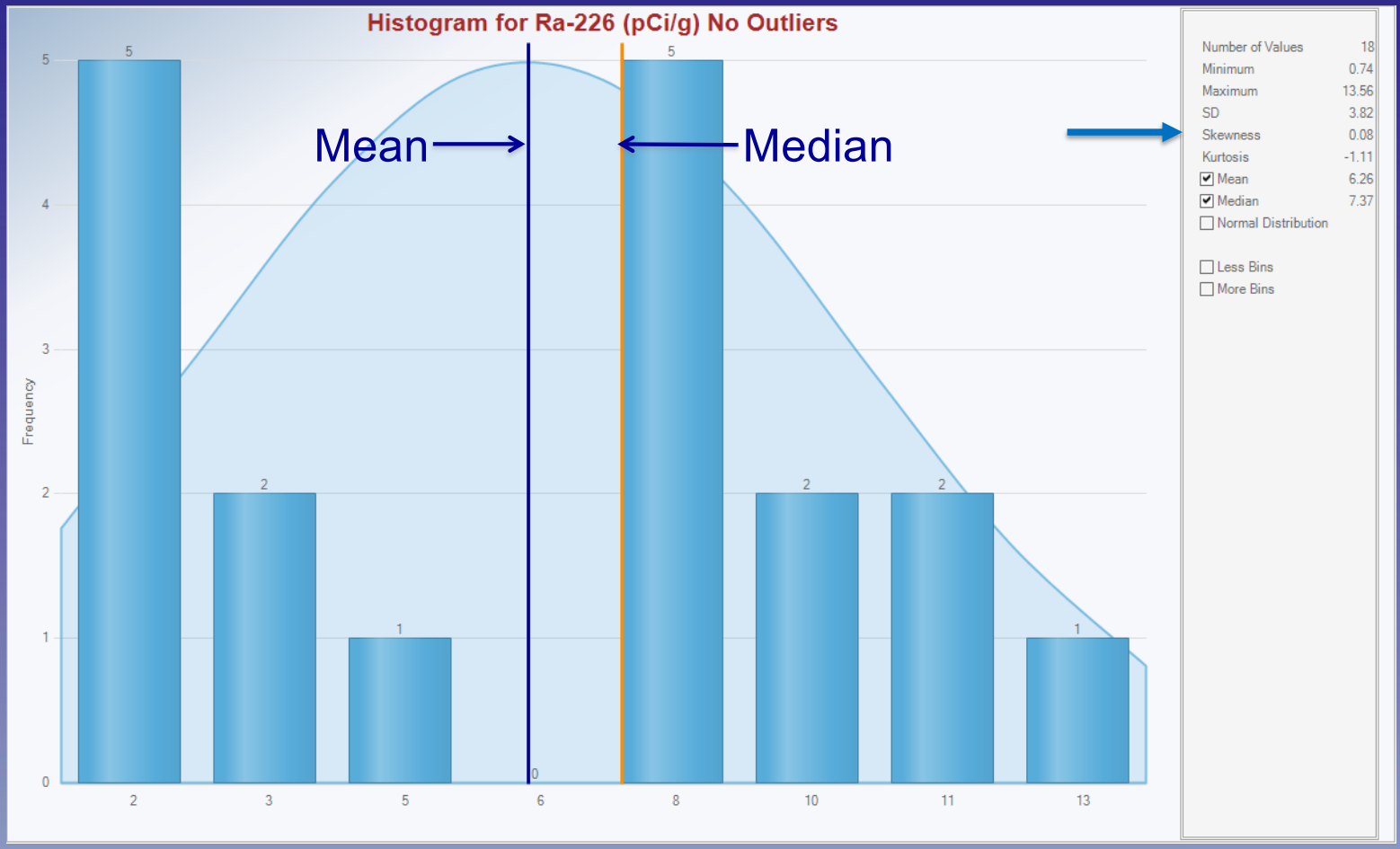
n = 20 → 18

Mean = 8.53 → 6.26

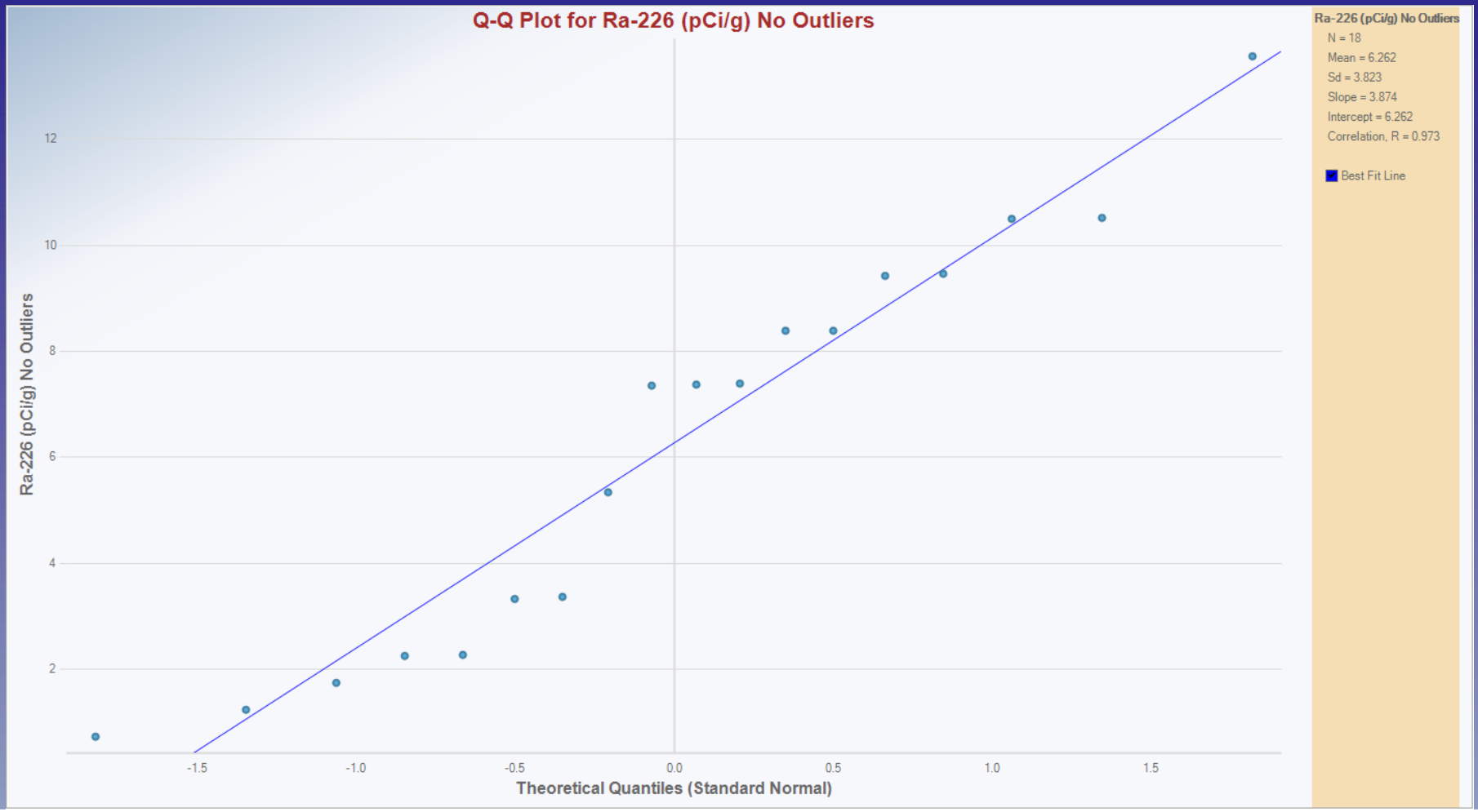
Median = 7.38 → 7.37

Standard Deviation = 8.18 → 3.82

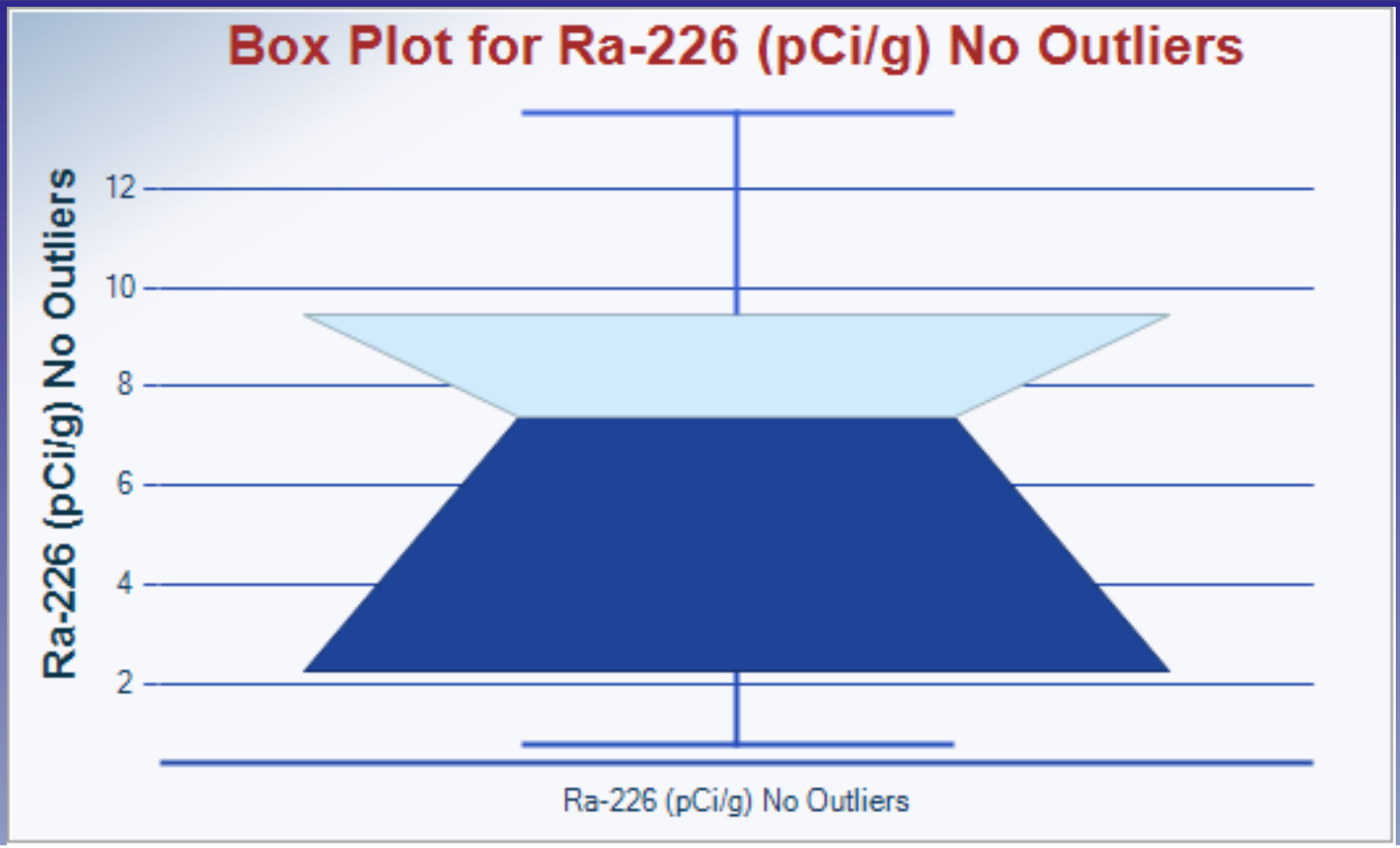
Histogram Without Outliers



Q-Q Plot Without Outliers

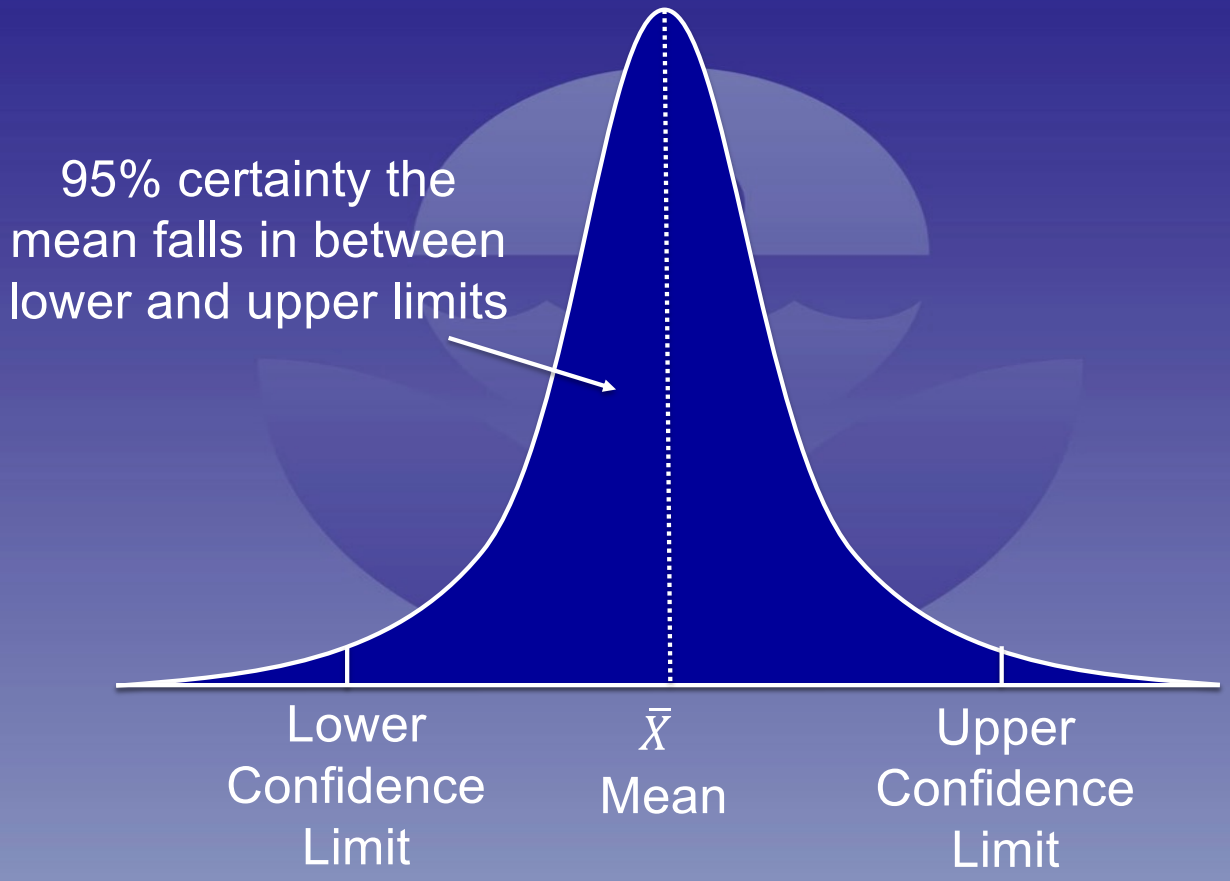


Box Plot Without Outliers



UCLs and BTVs

Confidence Interval



UCLs and BTVs

- Upper Confidence Limit (UCL): Upper limit of a confidence interval for a parameter of interest (typically the mean).
  - Frequently the exposure point concentration (EPC) in risk assessment
- Background Threshold Value (BTV): Upper value of background; a value greater than the BTV is considered contamination.
  - Upper Tolerance Limit (UTL)
  - Upper Prediction Limit (UPL)
  - Upper Simultaneous Limit (USL)

UCLs and BTVs With/Without Outliers

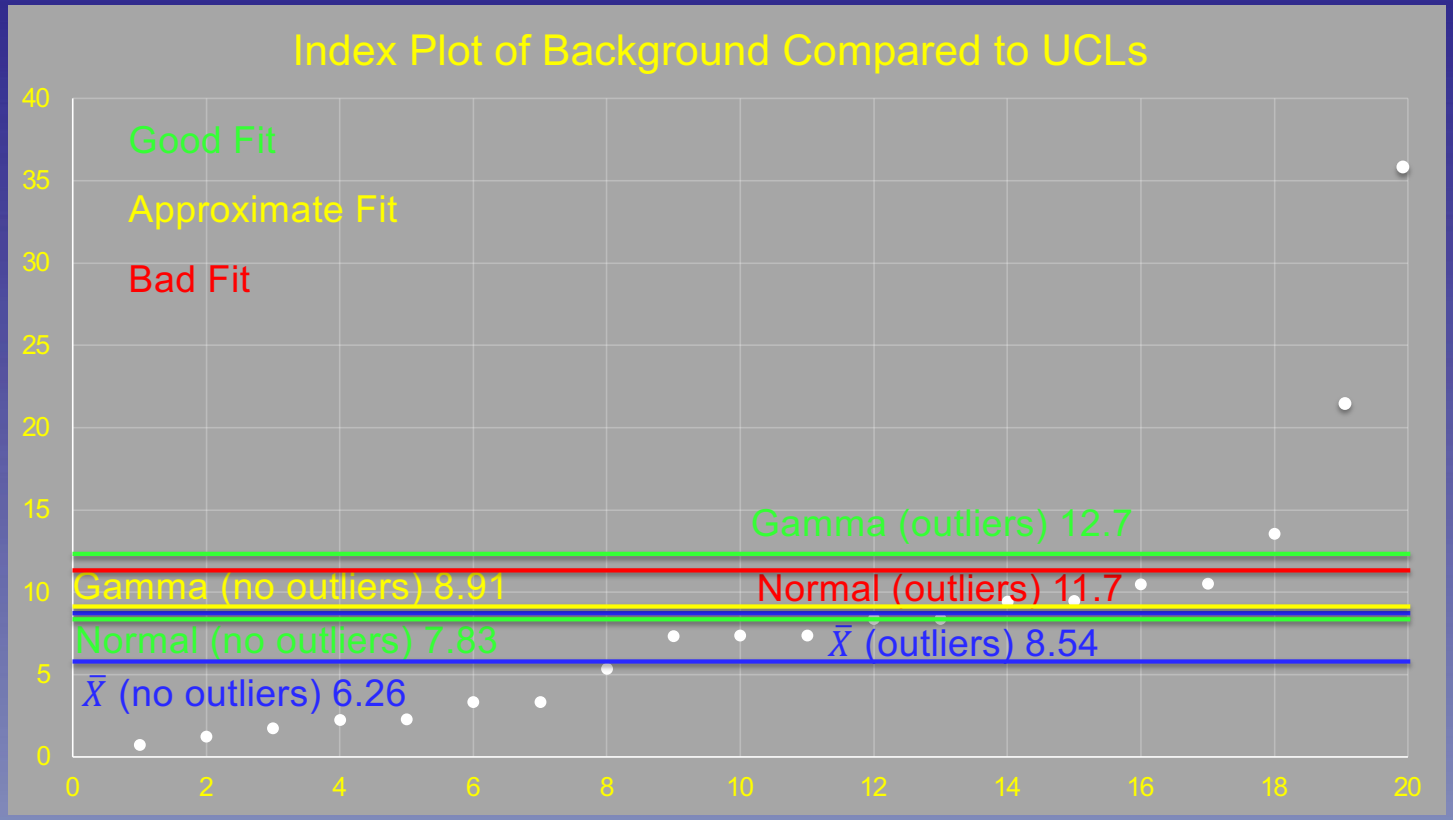
Distribution	95% UCL With Outliers	95% UCL Without Outliers	95% USL With Outliers	95% USL Without Outliers
	$\bar{X}$ = 8.54	$\bar{X}$ = 6.26	High = 35.9	High = 13.6
Gamma	12.7	8.91	38.4	23.2
Normal	11.7	7.83	29.5	15.8
~~Lognormal~~	~~18.3~~	~~13.0~~	~~68.6~~	~~39.9~~
Non-Parametric	11.6	7.74	35.9	13.6

Green = good fit

Yellow = approximate fit

Red = bad fit

Graphic Comparison of UCLs

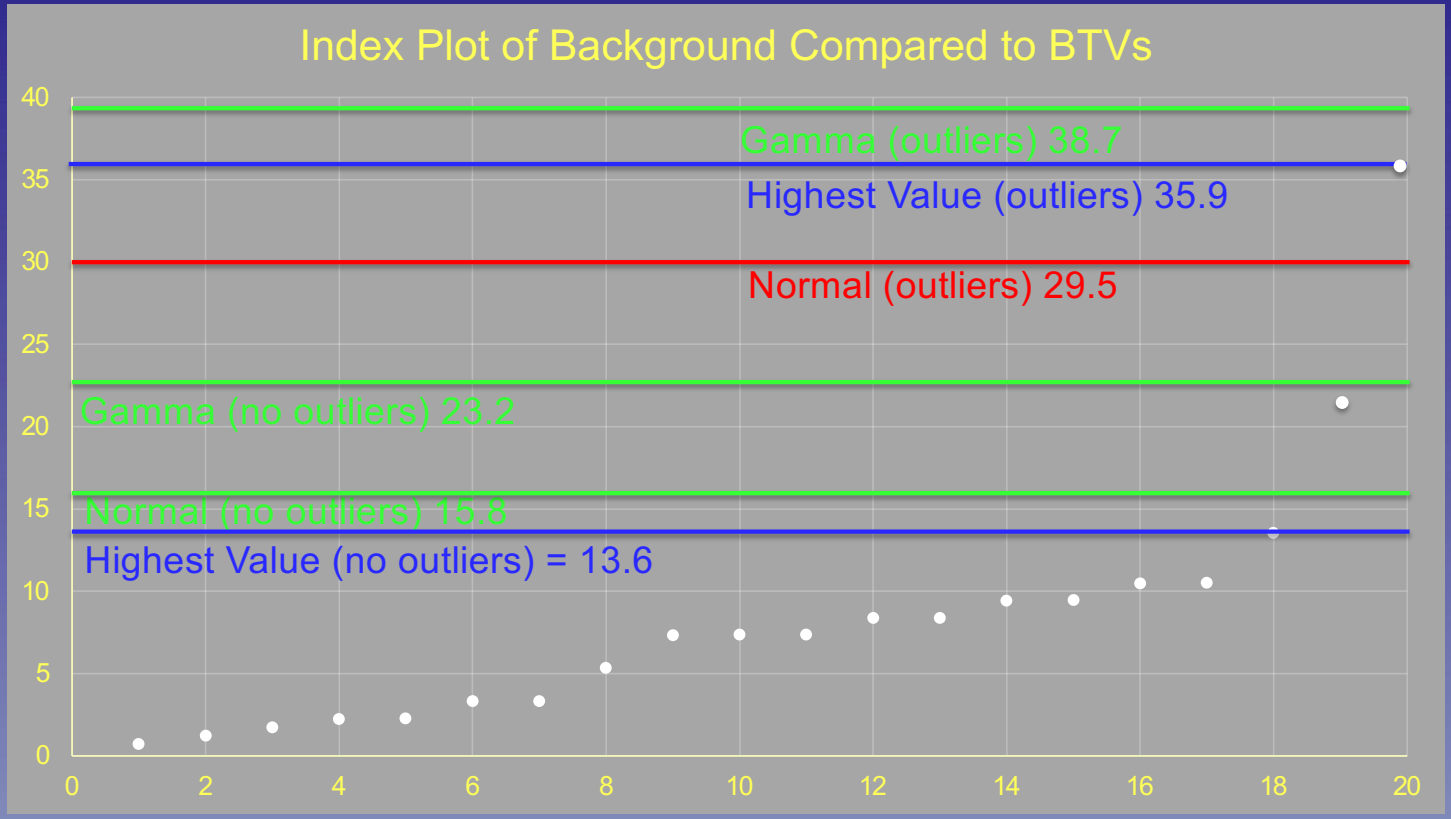


Graphic Comparison of BTVs

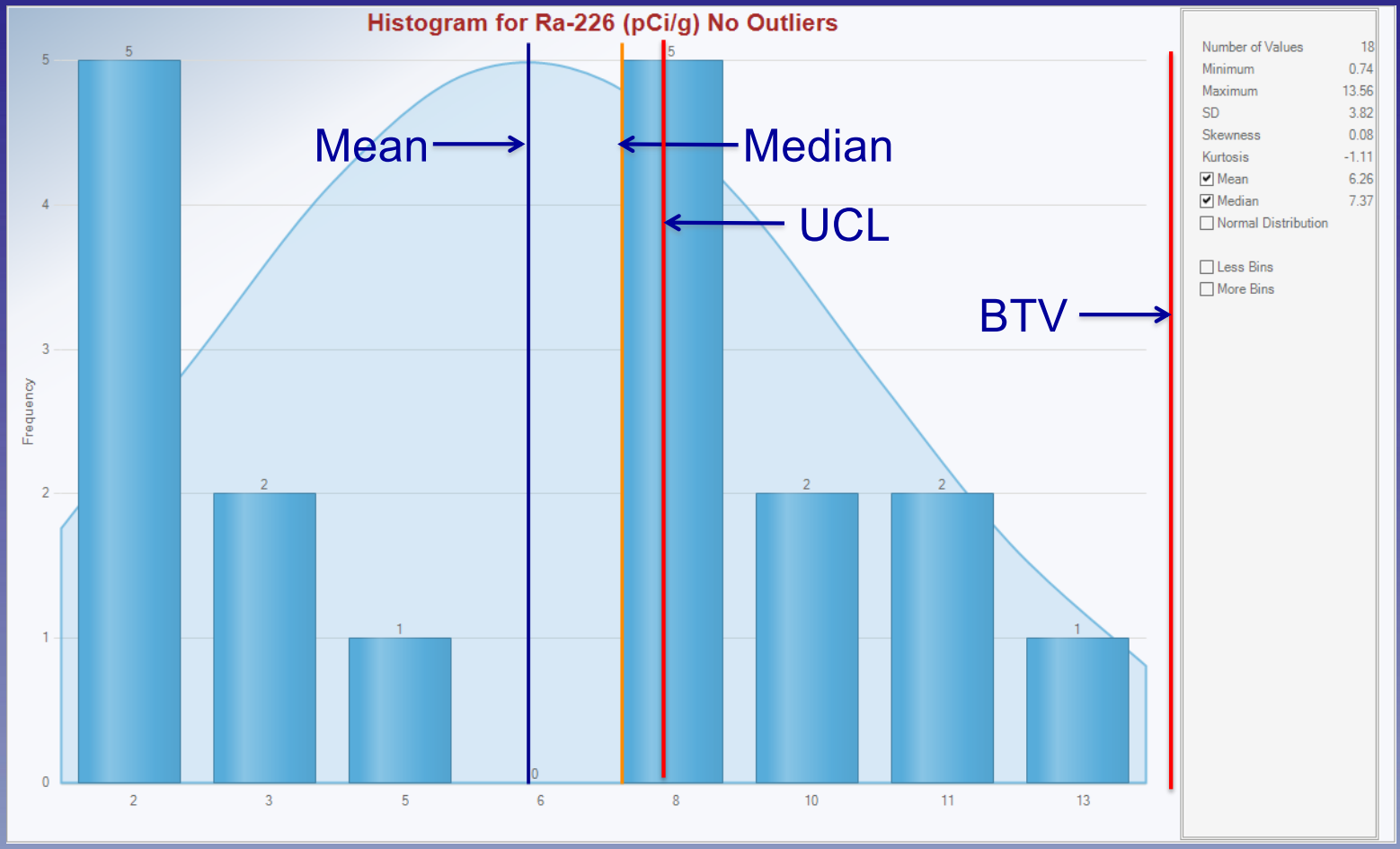
Good Fit

Approximate Fit

Bad Fit



Histogram Without Outliers

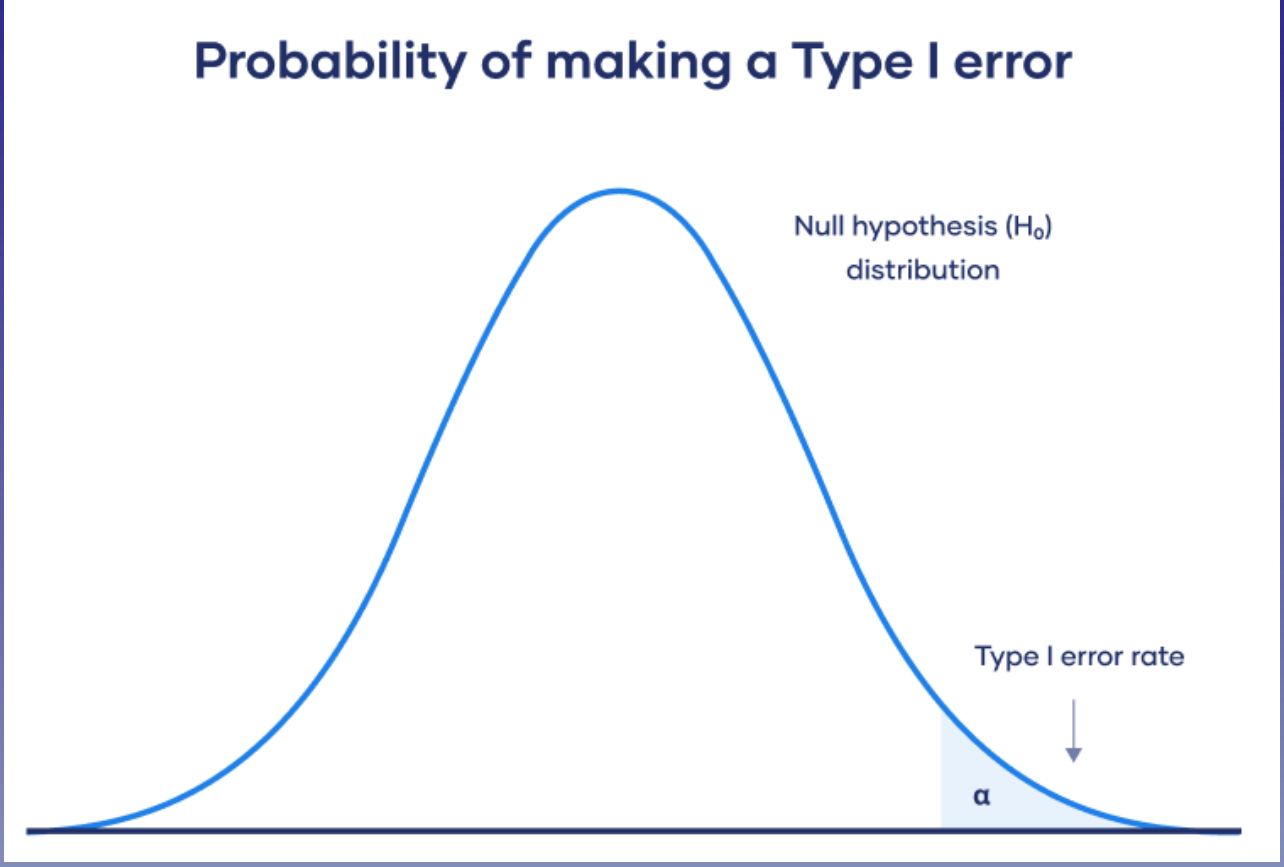


Hypothesis Testing & Decision Errors

Hypothesis Testing

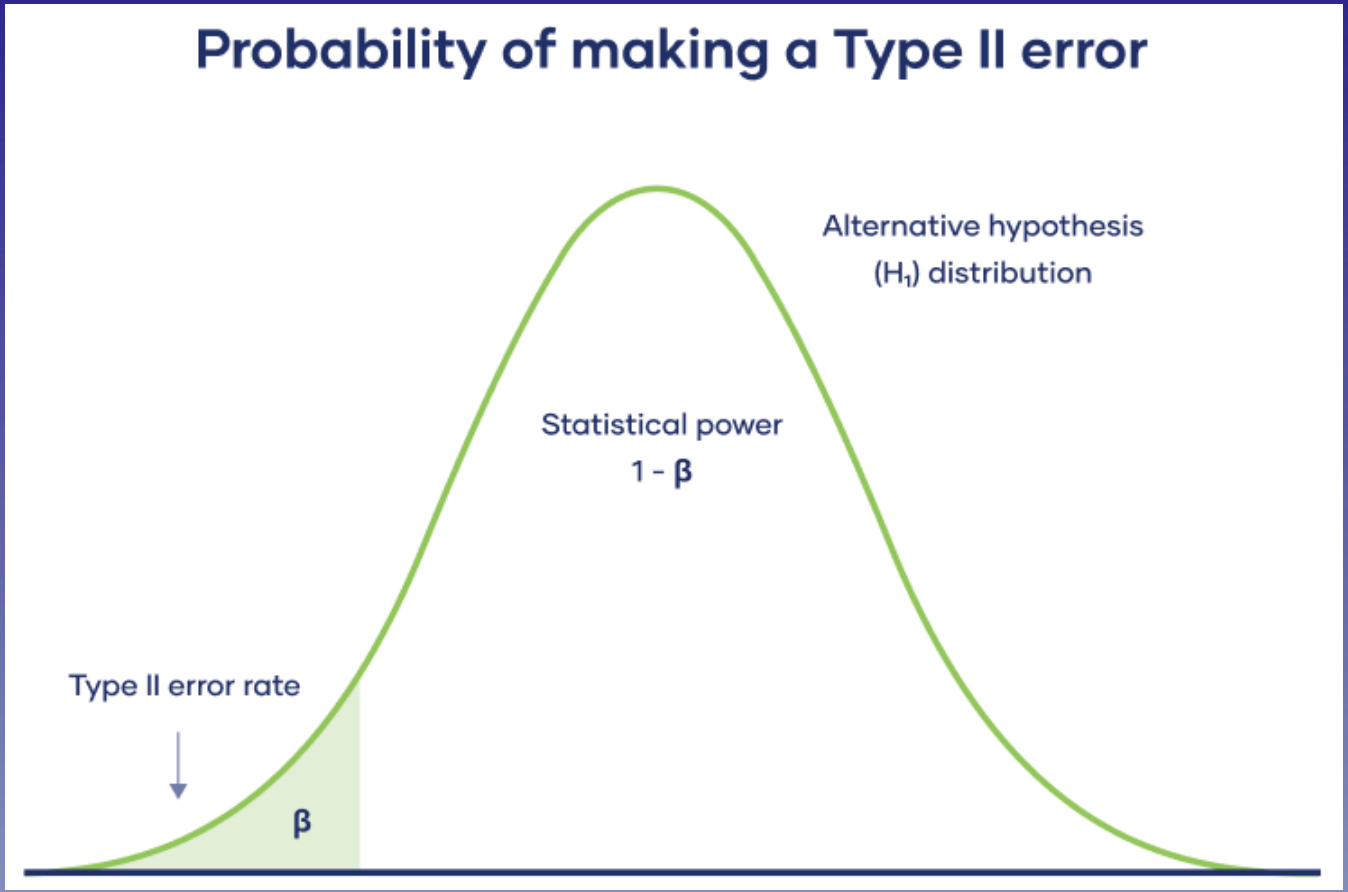
- Null hypothesis (Ho): The site is contaminated.
- Alternative hypothesis(Ha): The site is not contaminated.
- Decision errors:
  - Type I (alpha) – Ho is true (contaminated), but we decide false (not contaminated)
  - Type II (beta) – Ho is false (not contaminated), but we decide true (contaminated)
- Typically, we want to control Type I errors to protect human health and the environment

Alpha



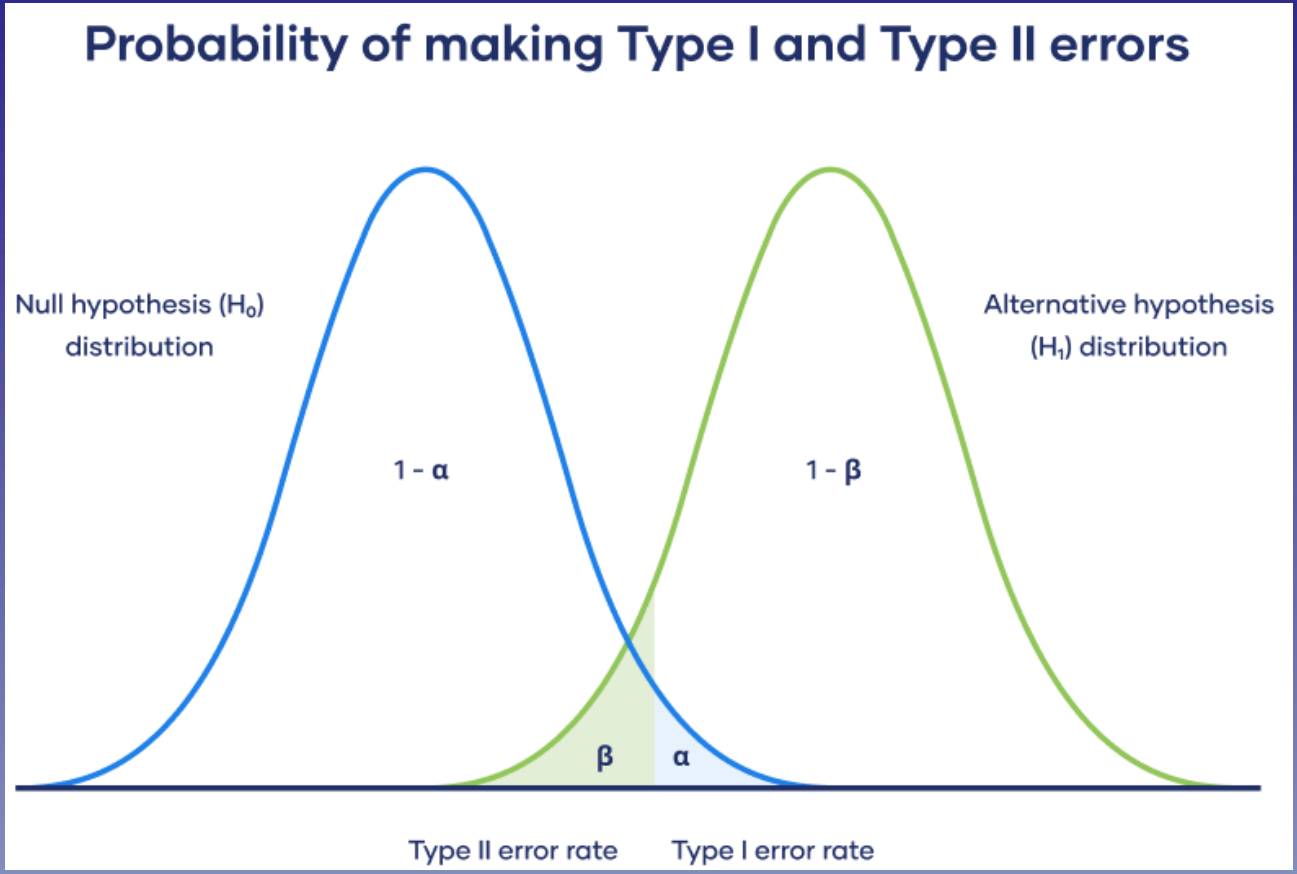
Bhandari, P. (2022, November 11). *Type I & Type II Errors | Differences, Examples, Visualizations.* Scribbr. Retrieved March 13, 2023, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Beta



Bhandari, P. (2022, November 11). *Type I & Type II Errors | Differences, Examples, Visualizations.* Scribbr. Retrieved March 13, 2023, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

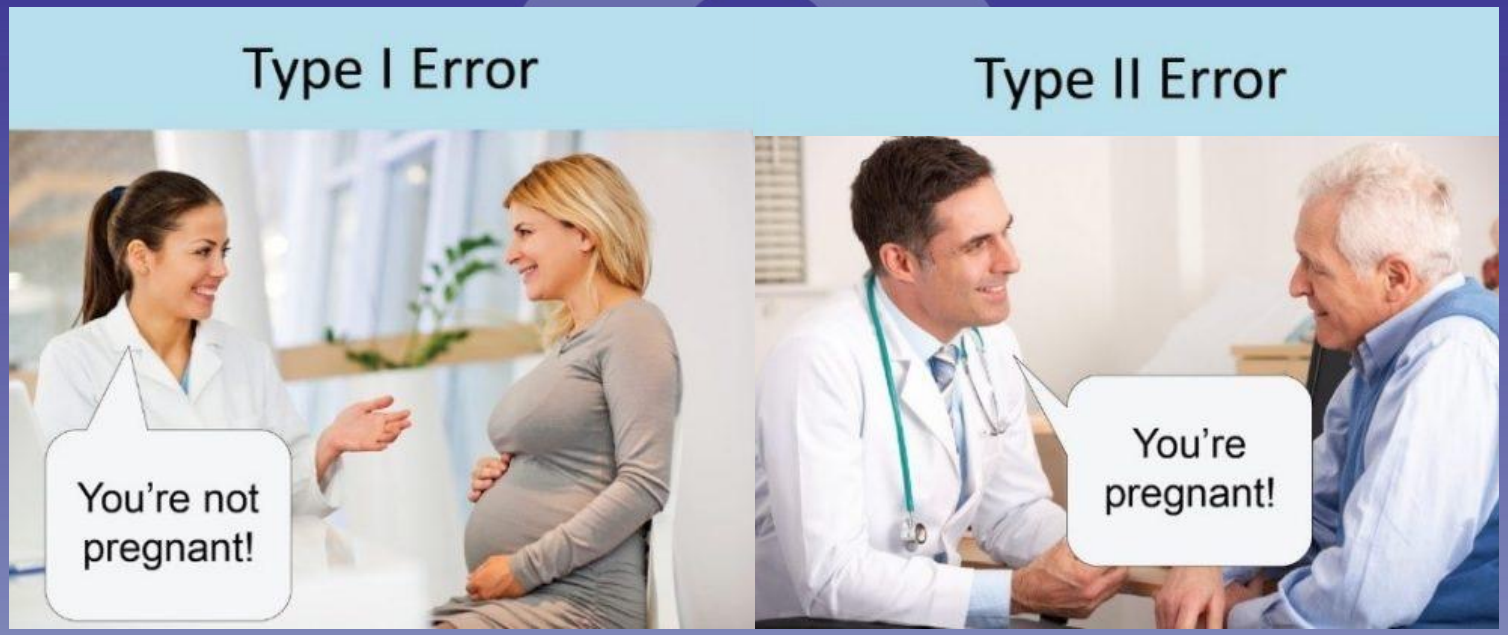
Type I and II Errors



Bhandari, P. (2022, November 11). *Type I & Type II Errors | Differences, Examples, Visualizations.* Scribbr. Retrieved March 13, 2023, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Decision Errors Can Be Serious!

Ho = You are pregnant



Adapted from: unbiasedresearch.blogspot.com

Decision Error Rates

- Alpha is set at 5% (typically)
- Beta is set at 10% (typically)

- How much data you need to collect to determine if Ho is true or false is dependent on alpha and beta (plus a few other variables)
  - $N = \frac{(Z_{1-\propto}+Z_{1-\beta})^2}{3(P_r-0.5)^2}$ (contamination found in background)
  - $N = \frac{(Z_{1-\propto}+Z_{1-\beta})^2}{4(Sign\ p\ -0.5)^2}$ (contamination not found in background)

Statistical Testing

- How do you know to accept or reject Ho (null hypothesis)?
- Wilcoxon Rank Sum (WRS) test (contamination found in background) compares two data sets
  - Non-parametric
- Sign test (contamination not found in background)
  - Non-parametric

Single Sample Compared to Action Level

- Frequently we want to decide clean vs contaminated based on a ***single*** sample

ONE SAMPLE!?

- Analytical accuracy – most analyses are 95% accurate

That's Pretty Good! – Right?

- Sampling accuracy – error in collecting representative sample is

5%?

100%?

50%?

350%?

2000%?

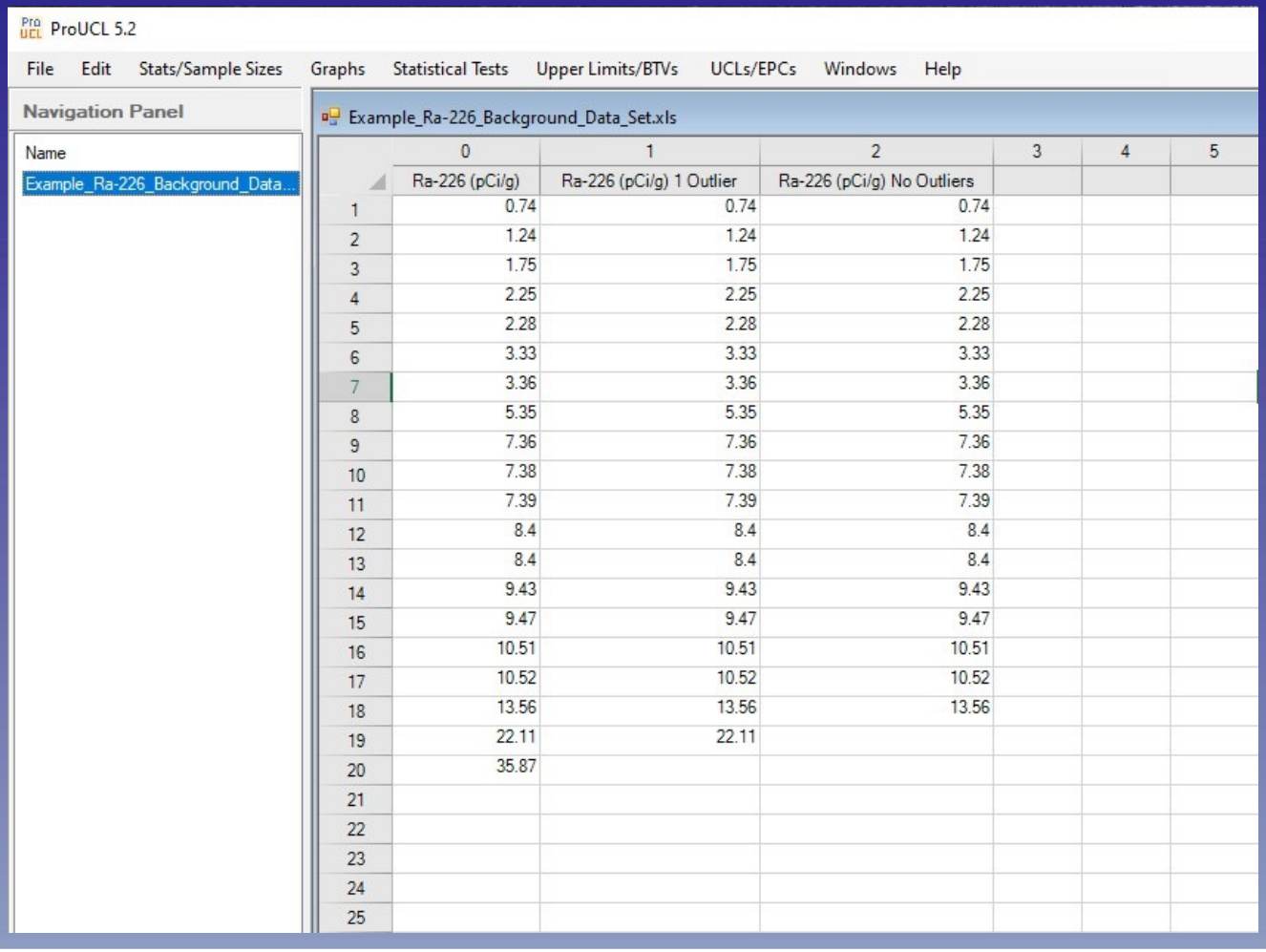
23%?

75%?

64%?

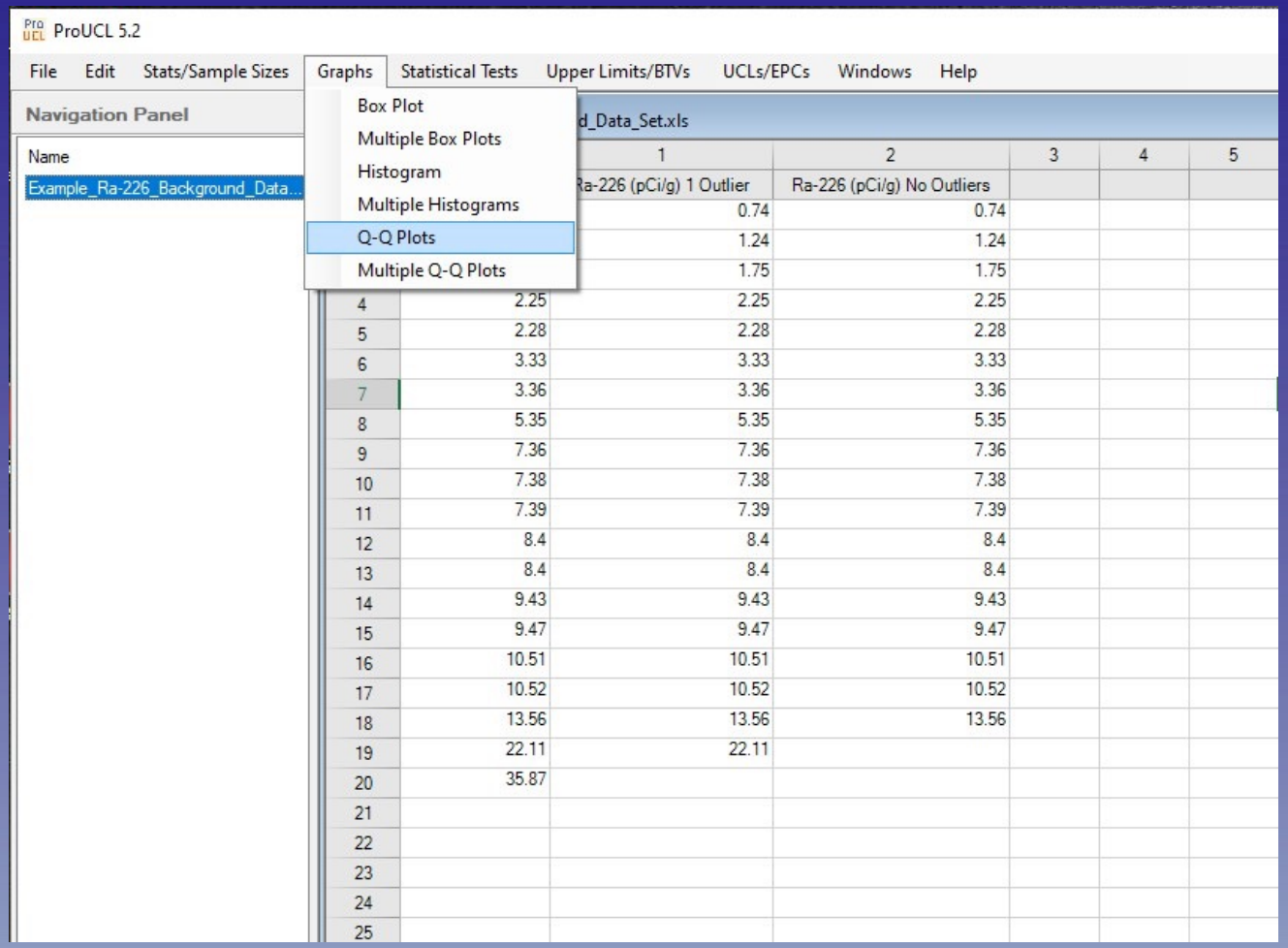
ProUCL
(Version 5.2)

ProUCL Menu

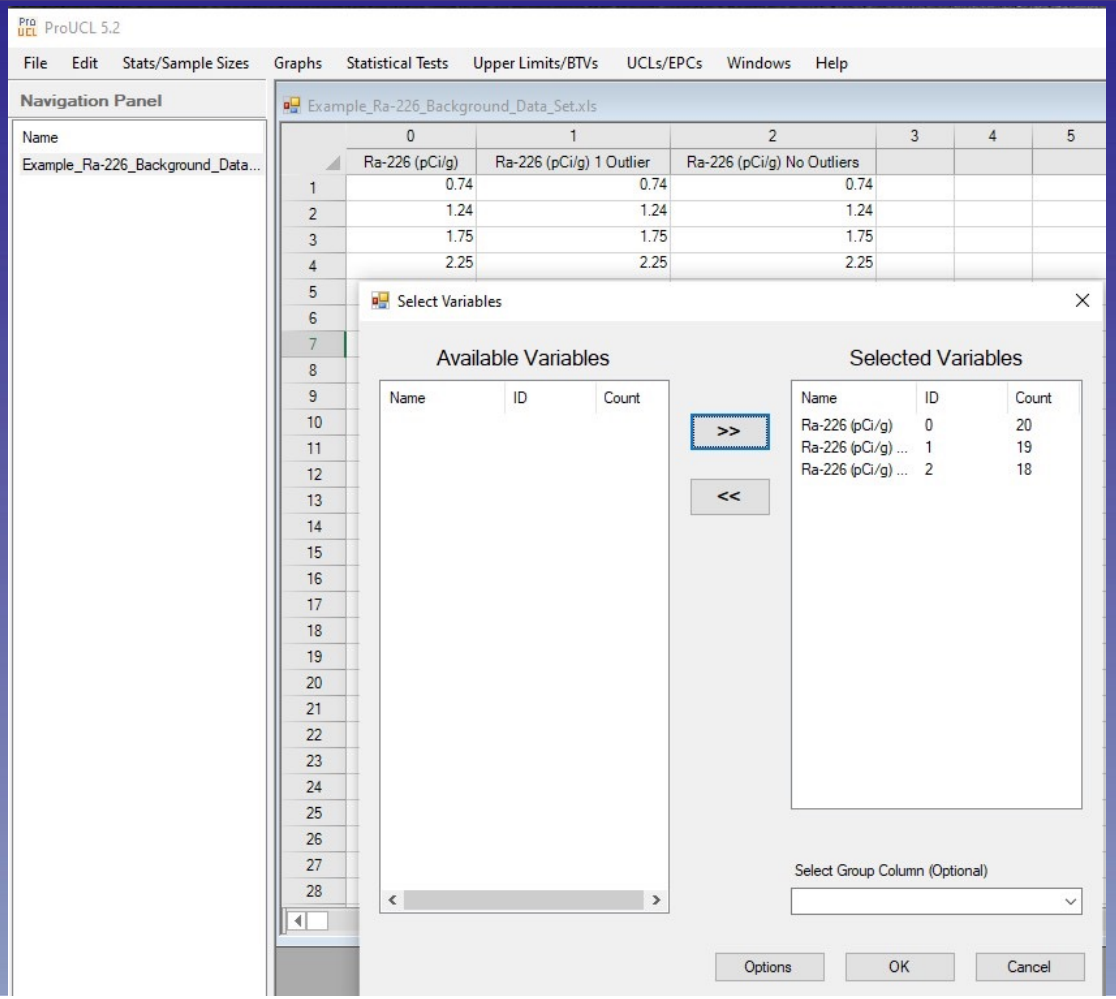


	0	1	2	3	4	5
	Ra-226 (pCi/g)	Ra-226 (pCi/g) 1 Outlier	Ra-226 (pCi/g) No Outliers			
1	0.74	0.74	0.74			
2	1.24	1.24	1.24			
3	1.75	1.75	1.75			
4	2.25	2.25	2.25			
5	2.28	2.28	2.28			
6	3.33	3.33	3.33			
7	3.36	3.36	3.36			
8	5.35	5.35	5.35			
9	7.36	7.36	7.36			
10	7.38	7.38	7.38			
11	7.39	7.39	7.39			
12	8.4	8.4	8.4			
13	8.4	8.4	8.4			
14	9.43	9.43	9.43			
15	9.47	9.47	9.47			
16	10.51	10.51	10.51			
17	10.52	10.52	10.52			
18	13.56	13.56	13.56			
19	22.11	22.11				
20	35.87					
21						
22						
23						
24						
25						

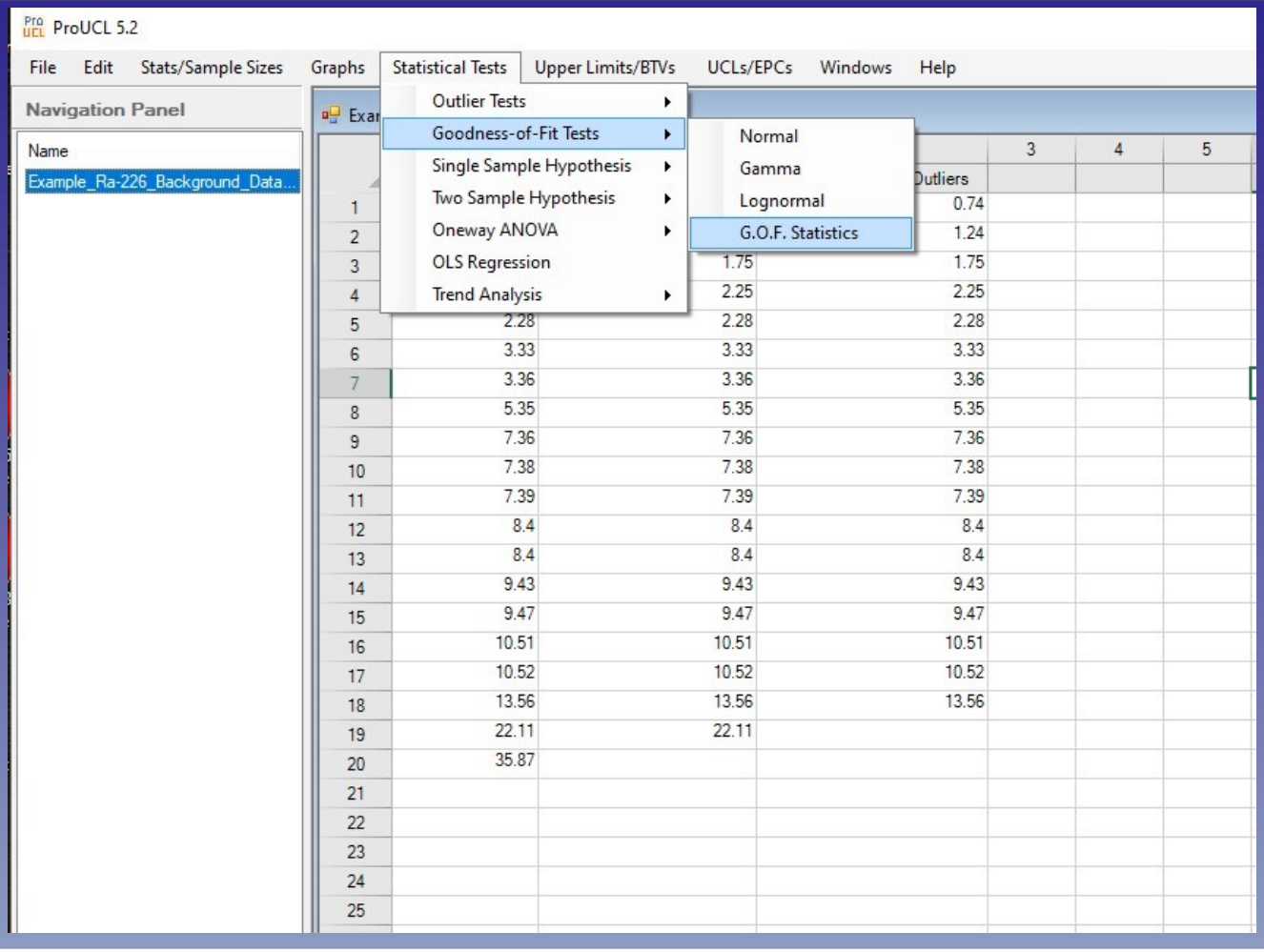
Graphs


ProUCL 5.2
File
Edit
Stats/Sample Sizes
Graphs
Statistical Tests
Upper Limits/BTVs
UCLs/EPCs
Windows
Help
Box Plot
Multiple Box Plots
Histogram
Multiple Histograms
Q-Q Plots
Multiple Q-Q Plots
Navigation Panel
Name
Example_Ra-226_Background_Data...
d_Data_Set.xls
1
2
3
4
5
Ra-226 (pCi/g) 1 Outlier
Ra-226 (pCi/g) No Outliers
0.74 0.74
1.24 1.24
1.75 1.75
2.25 2.25 2.25
2.28 2.28 2.28
3.33 3.33 3.33
3.36 3.36 3.36
5.35 5.35 5.35
7.36 7.36 7.36
7.38 7.38 7.38
7.39 7.39 7.39
8.4 8.4 8.4
8.4 8.4 8.4
9.43 9.43 9.43
9.47 9.47 9.47
10.51 10.51 10.51
10.52 10.52 10.52
13.56 13.56 13.56
22.11 22.11
35.87

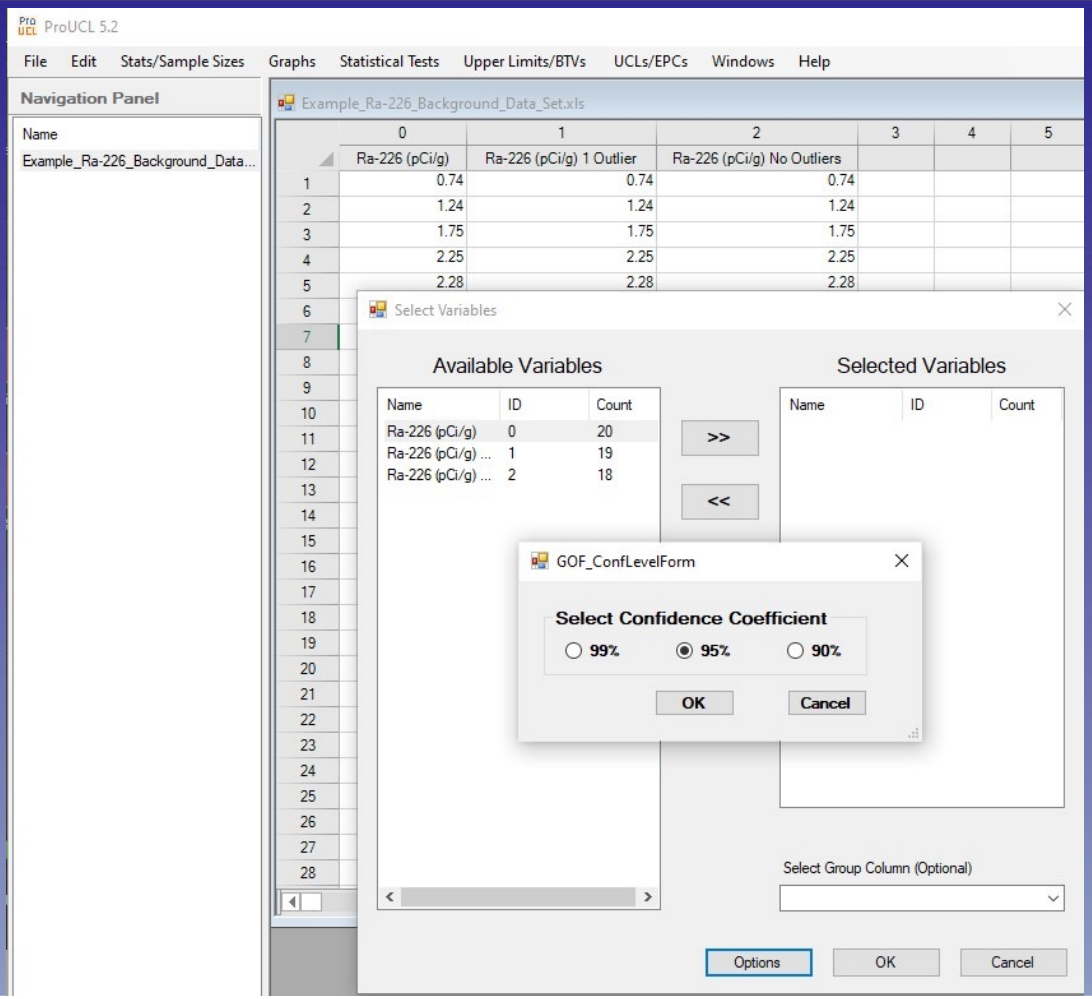
Option to Select Multiple Data Sets



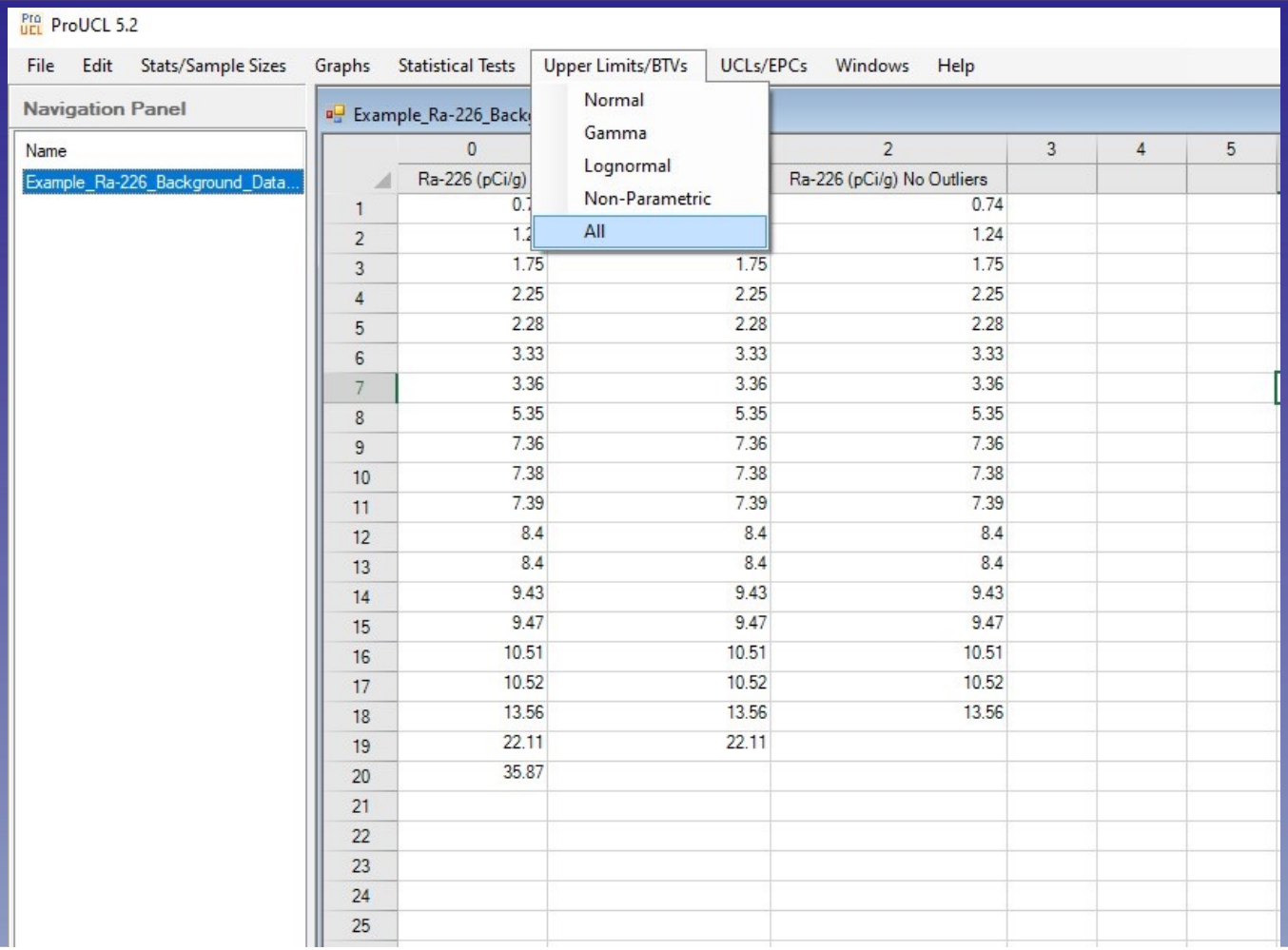
Statistical Tests



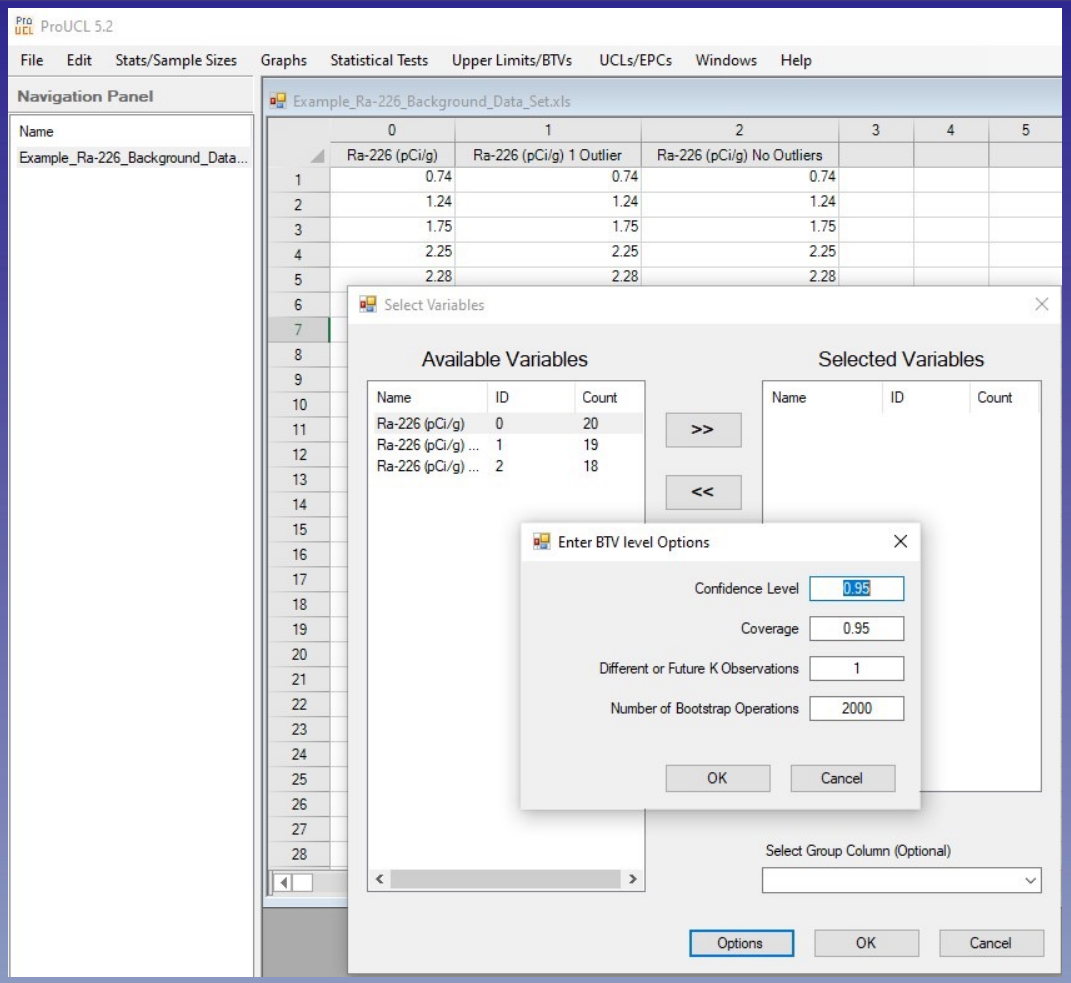
“Options” Can Be Changed



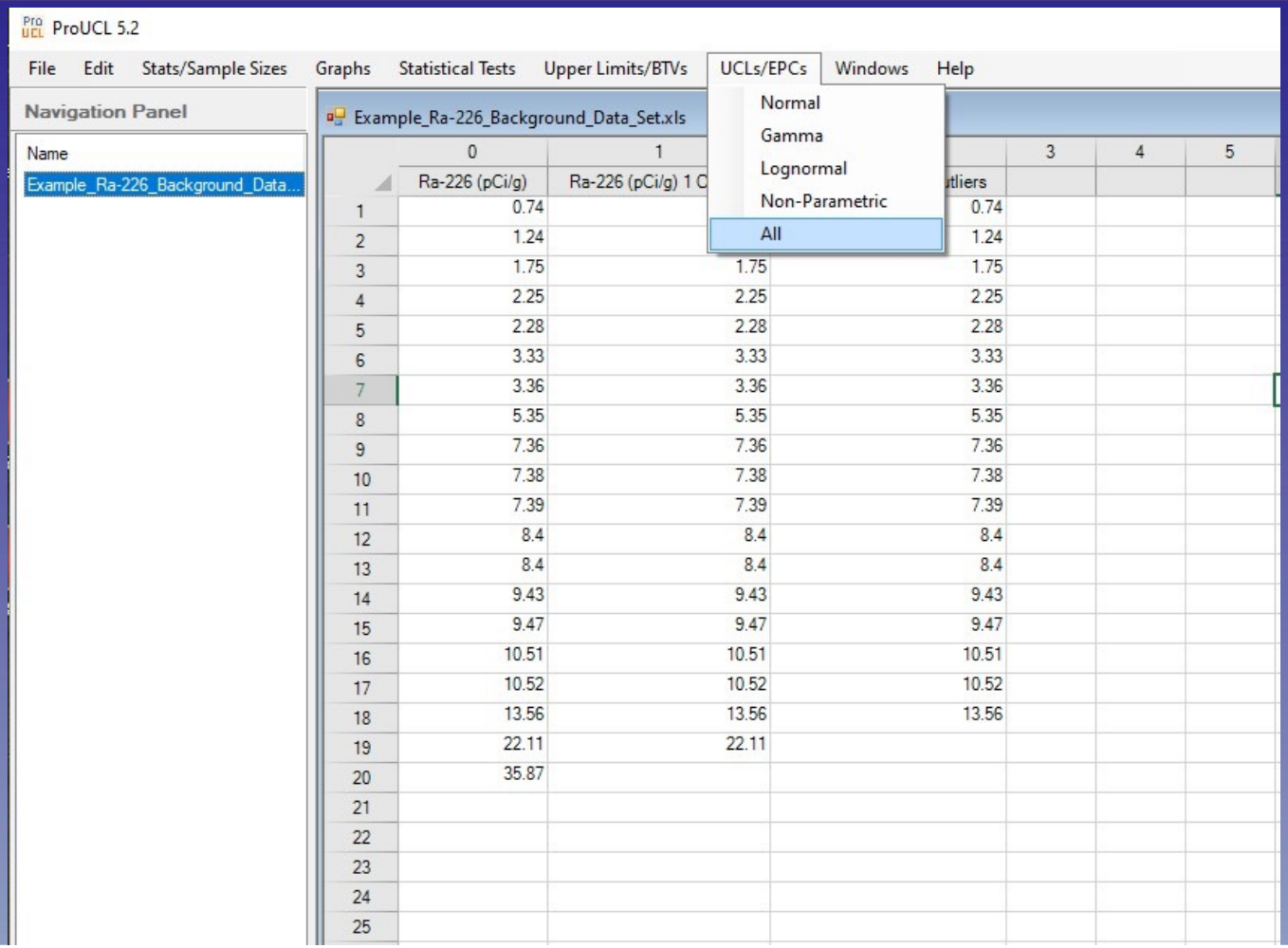
BTV Menu


ProUCL 5.2
File
Edit
Stats/Sample Sizes
Graphs
Statistical Tests
Upper Limits/BTVs
UCLs/EPCs
Windows
Help
Normal
Gamma
Lognormal
Non-Parametric
All
Navigation Panel
Name
Example_Ra-226_Background_Data...
Example_Ra-226_Back
Ra-226 (pCi/g)
Ra-226 (pCi/g) No Outliers

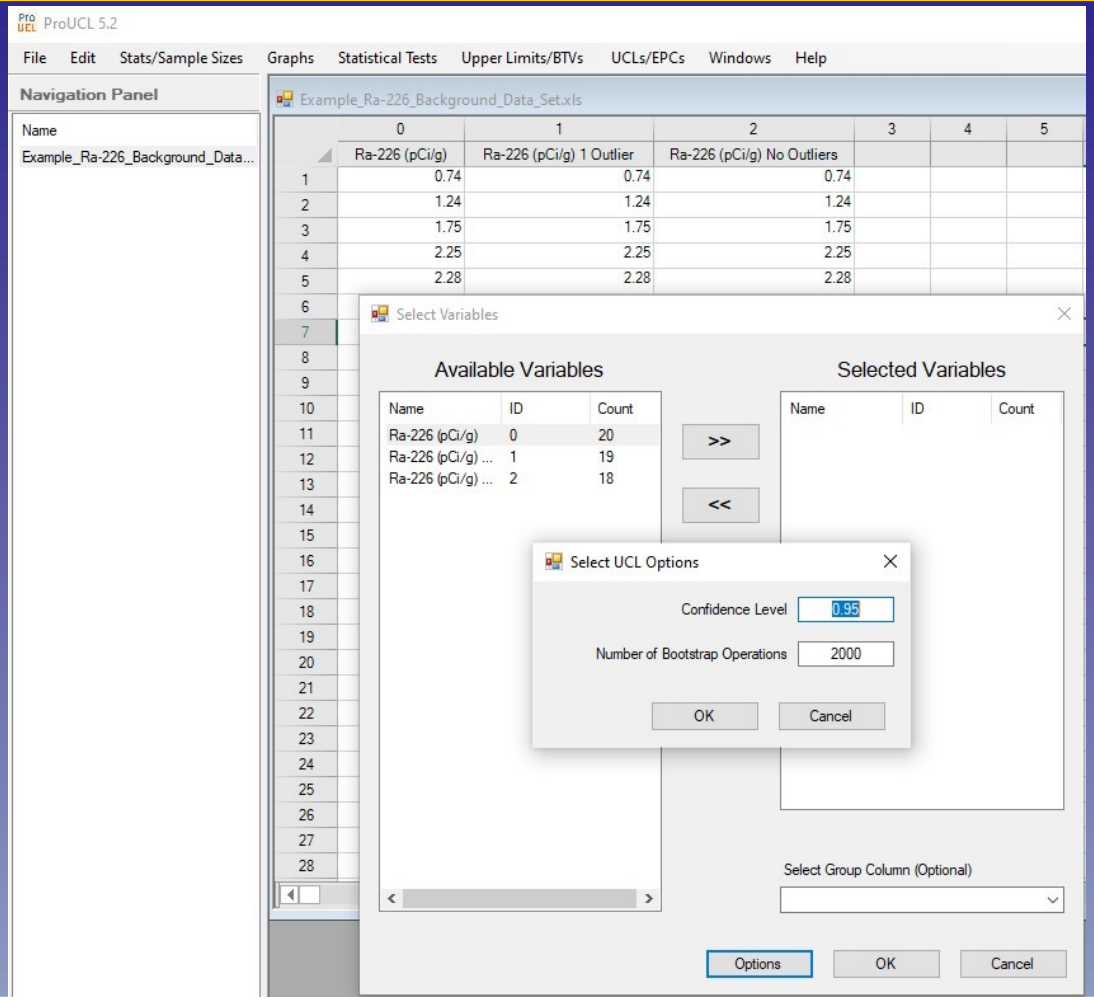
BTV Options


ProUCL 5.2
File
Edit
Stats/Sample Sizes
Graphs
Statistical Tests
Upper Limits/BTVs
UCLs/EPCs
Windows
Help
Navigation Panel
Name
Example_Ra-226_Background_Data...
Example_Ra-226_Background_Data_Set.xls
Ra-226 (pCi/g)
Ra-226 (pCi/g) 1 Outlier
Ra-226 (pCi/g) No Outliers
Select Variables
Available Variables
Selected Variables
Name
ID
Count
Ra-226 (pCi/g) 0 20
Ra-226 (pCi/g) ... 1 19
Ra-226 (pCi/g) ... 2 18
>>
<<
Enter BTV level Options
Confidence Level 0.95
Coverage 0.95
Different or Future K Observations 1
Number of Bootstrap Operations 2000
OK
Cancel
Select Group Column (Optional)
Options
OK
Cancel

UCL Menu


ProUCL 5.2
File
Edit
Stats/Sample Sizes
Graphs
Statistical Tests
Upper Limits/BTVs
UCLs/EPCs
Windows
Help
Normal
Gamma
Lognormal
Non-Parametric
All
Navigation Panel
Name
Example_Ra-226_Background_Data...
Example_Ra-226_Background_Data_Set.xls
Ra-226 (pCi/g)
Ra-226 (pCi/g) 1 C
tliers

UCL Options


ProUCL 5.2
File
Edit
Stats/Sample Sizes
Graphs
Statistical Tests
Upper Limits/BTVs
UCLs/EPCs
Windows
Help
Navigation Panel
Name
Example_Ra-226_Background_Data...
Example_Ra-226_Background_Data_Set.xls
Ra-226 (pCi/g)
Ra-226 (pCi/g) 1 Outlier
Ra-226 (pCi/g) No Outliers
0.74
1.24
1.75
2.25
2.28
Select Variables
Available Variables
Selected Variables
Name
ID
Count
Ra-226 (pCi/g) 0 20
Ra-226 (pCi/g) ... 1 19
Ra-226 (pCi/g) ... 2 18
>>
<<
Select UCL Options
Confidence Level
0.95
Number of Bootstrap Operations
2000
OK
Cancel
Select Group Column (Optional)
Options
OK
Cancel

Questions?

Carl Palladino
Health Physicist
415-336-1556
carl@palladinocompany.com