

# Statistical Analysis of Environmental Data

Presented by:

Carl Palladino

The Palladino Company, Inc.

Presented for:

Environmental Protection Agency, Region 9

March 21, 2023

# Instructor

---

**Carl Palladino**

Health Physicist

415-336-1556

[carl@palladinocompany.com](mailto:carl@palladinocompany.com)



# Course Sponsor

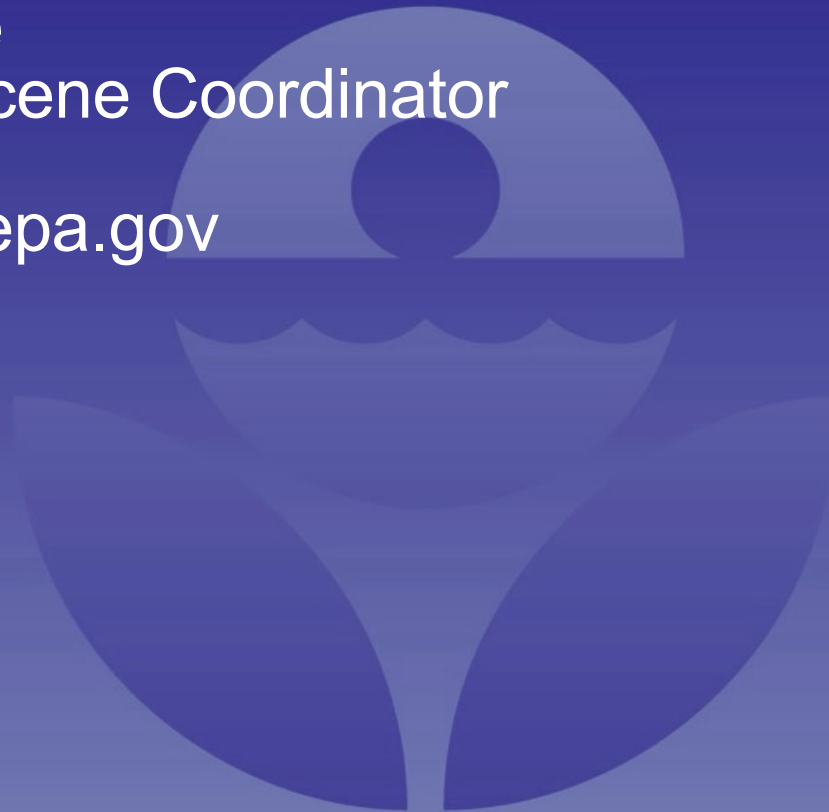
---

**Robert Wise**

Federal On-Scene Coordinator

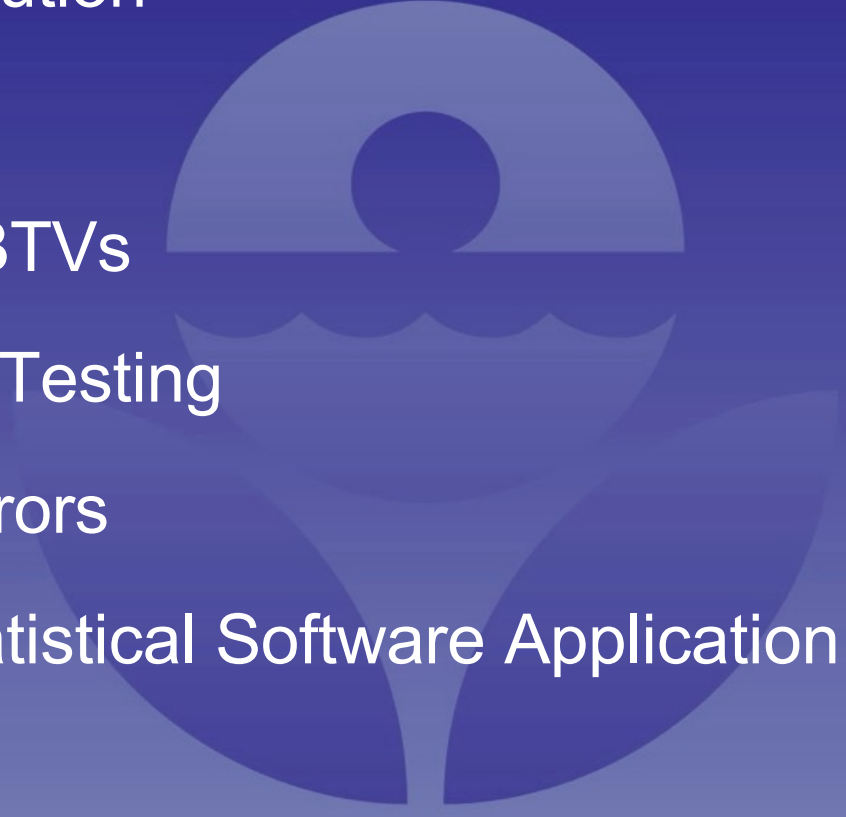
562-889-2572

[wise.robert@epa.gov](mailto:wise.robert@epa.gov)



# Course Agenda

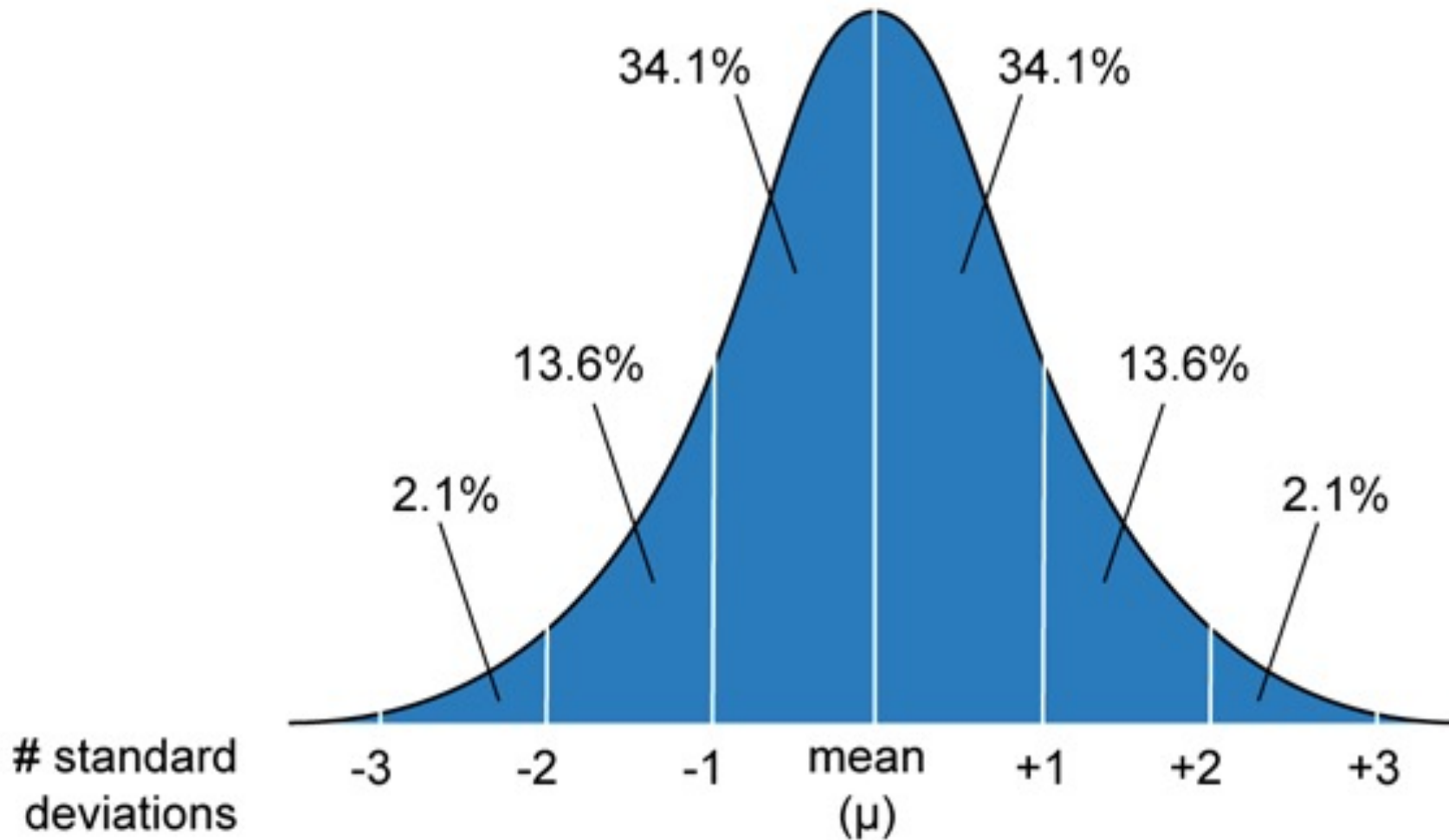
---

1. Data Distribution
  2. Outliers
  3. UCLs and BTVs
  4. Hypothesis Testing
  5. Decision Errors
  6. ProUCL Statistical Software Application
- 



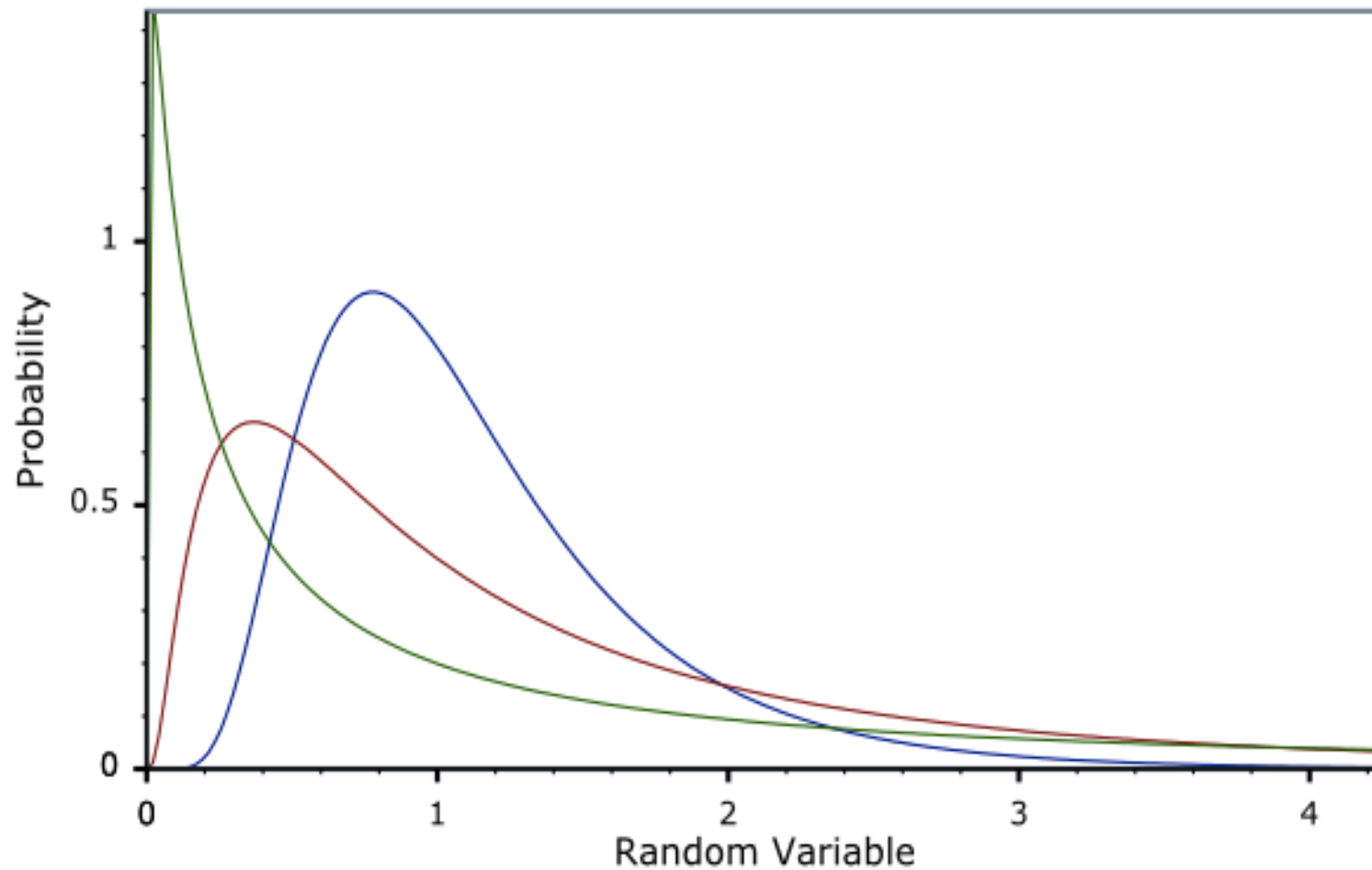
# Data Distribution

# Symmetric Normal Distribution

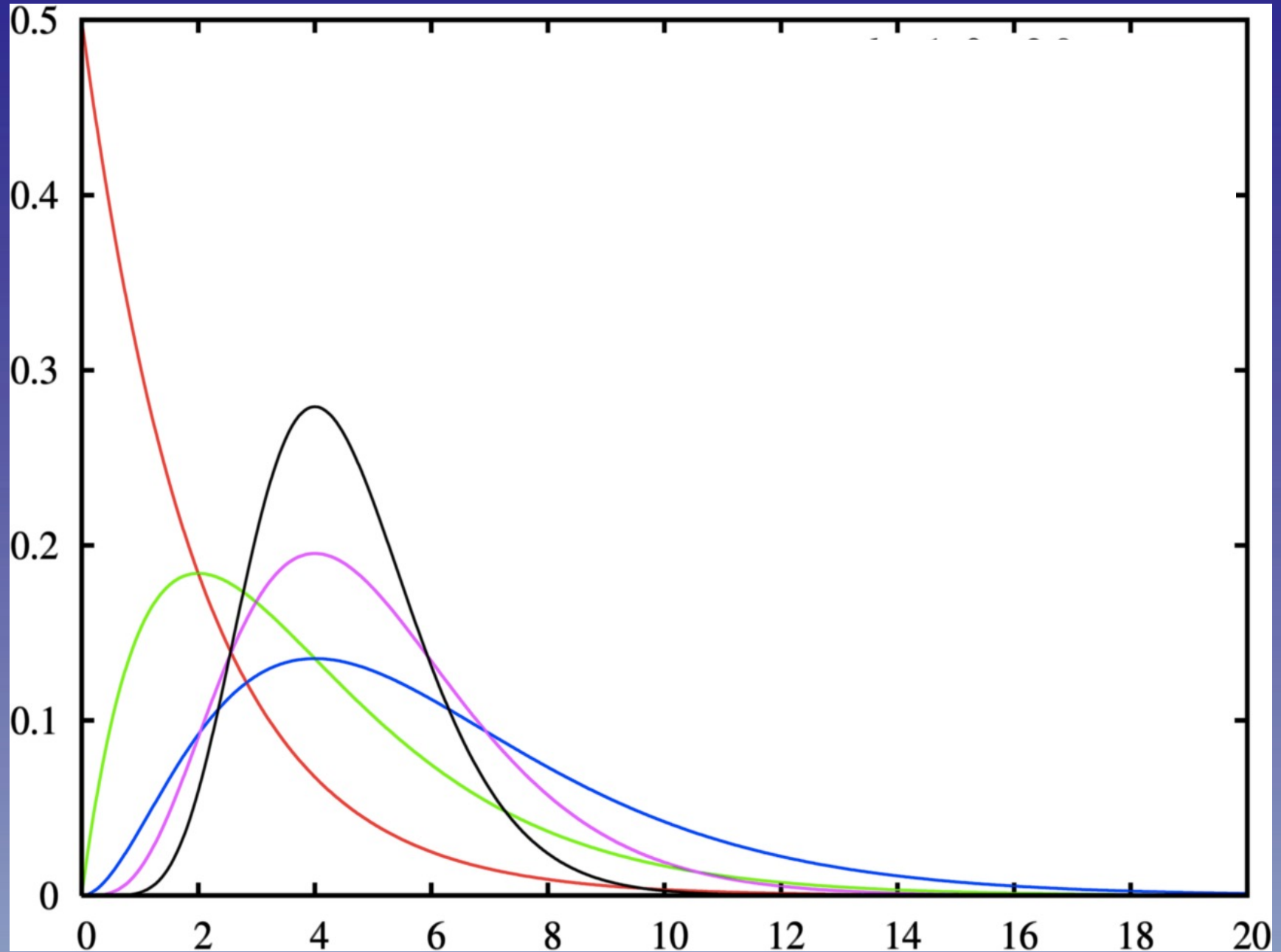


# Lognormal Distributions

---



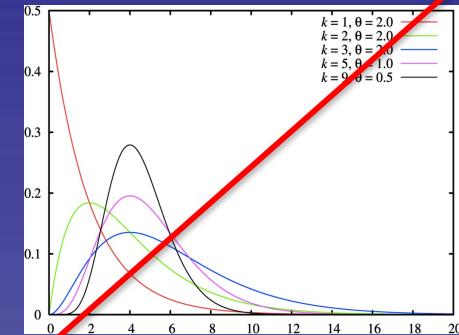
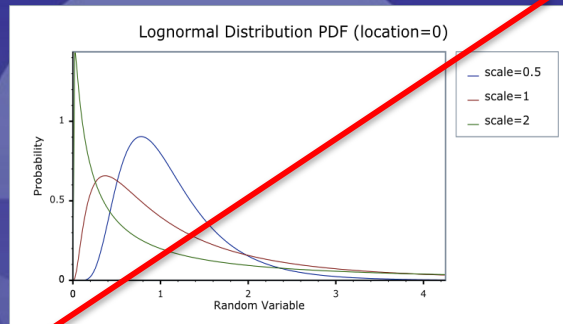
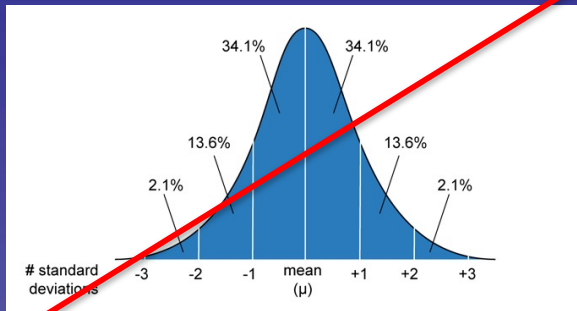
# Gamma Distributions





# Nonparametric Distribution

- Data does not fit a normal distribution



- Nonparametric statistics do not assume predefined distribution parameters
- Downside to nonparametric statistics is reduced power

# Example Data Set

---

Ra-226 Background Data Set of Size 20

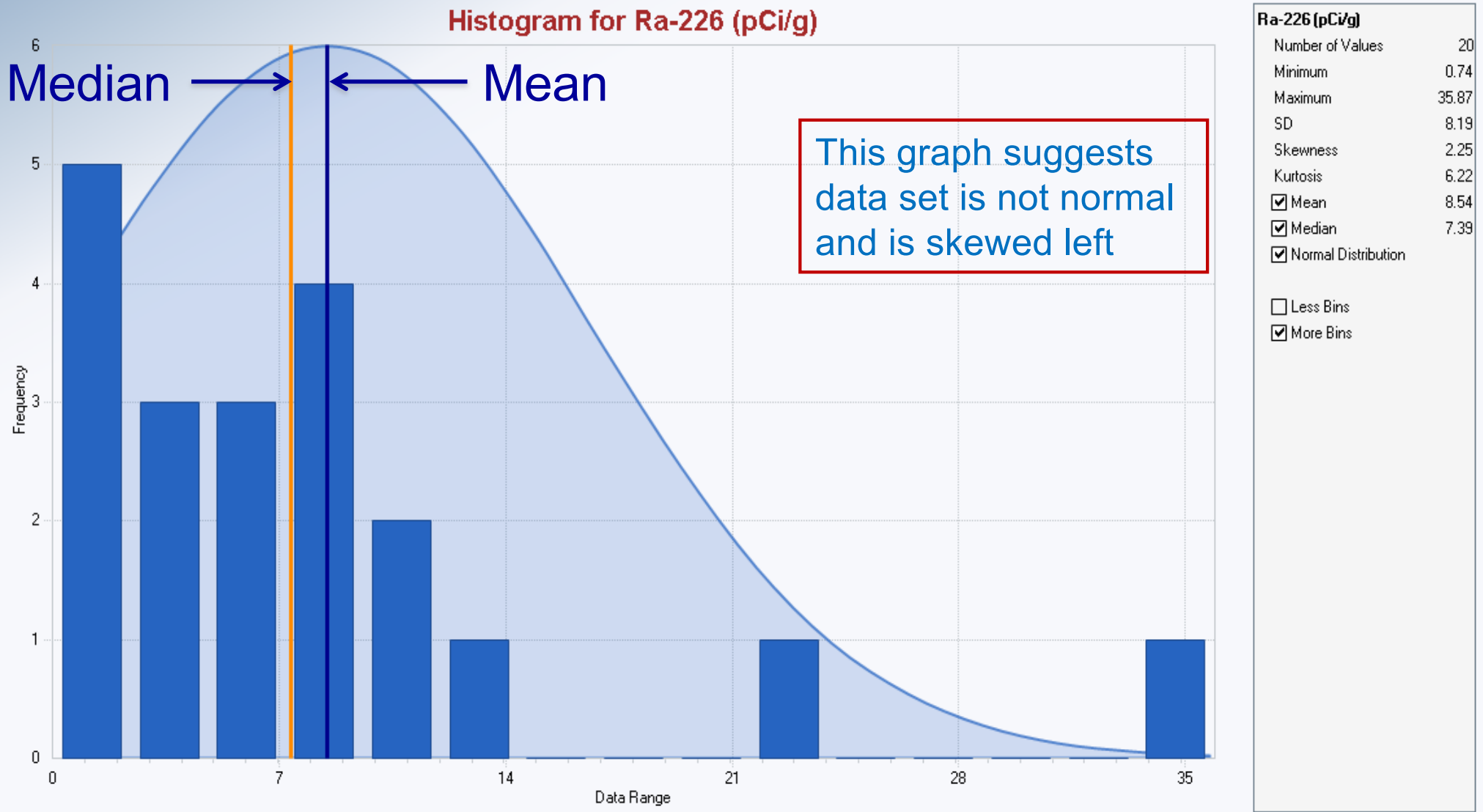
1)	0.740	11)	7.39
2)	1.24	12)	8.40
3)	1.75	13)	8.40
4)	2.25	14)	9.43
5)	2.28	15)	9.47
6)	3.33	16)	10.51
7)	3.36	17)	10.52
8)	5.35	18)	13.56
9)	7.36	19)	22.11
10)	7.38	20)	35.87

Mean = 8.53

Median = 7.38

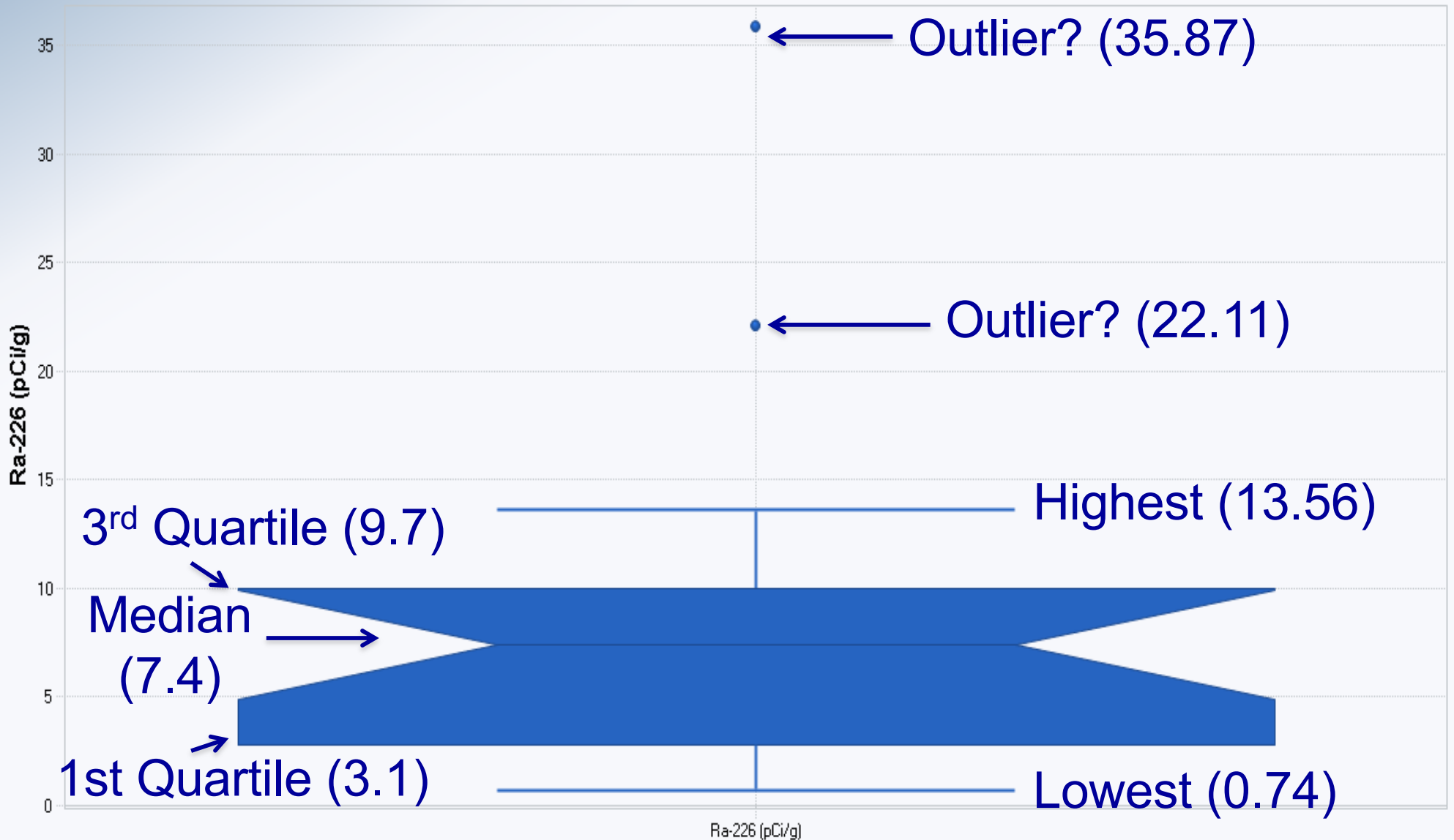
Standard  
Deviation = 8.18

# Histogram

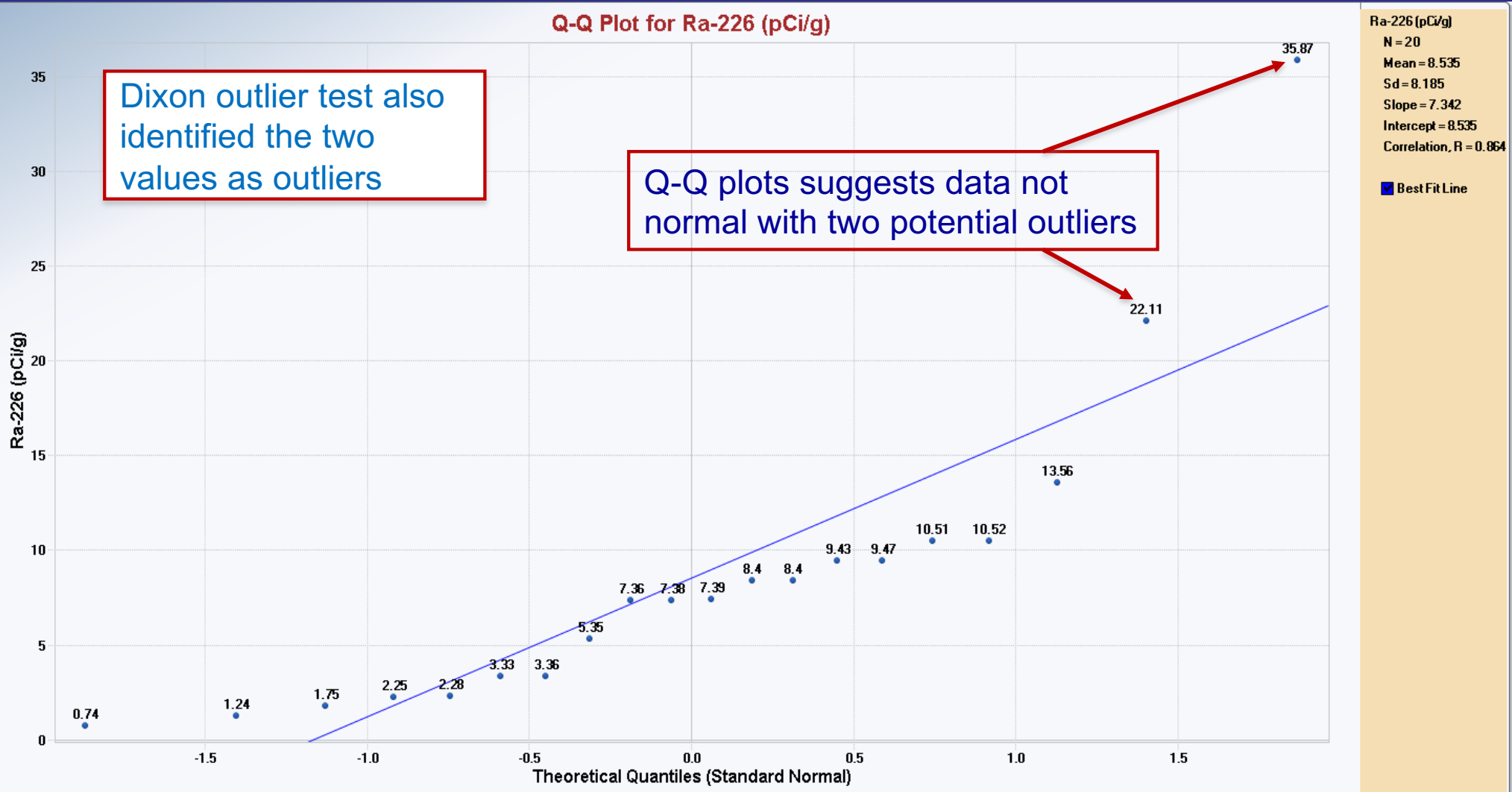


# Box Plot

Box Plot for Ra-226 (pCi/g)



# Quantile-Quantile (Q-Q) Plot

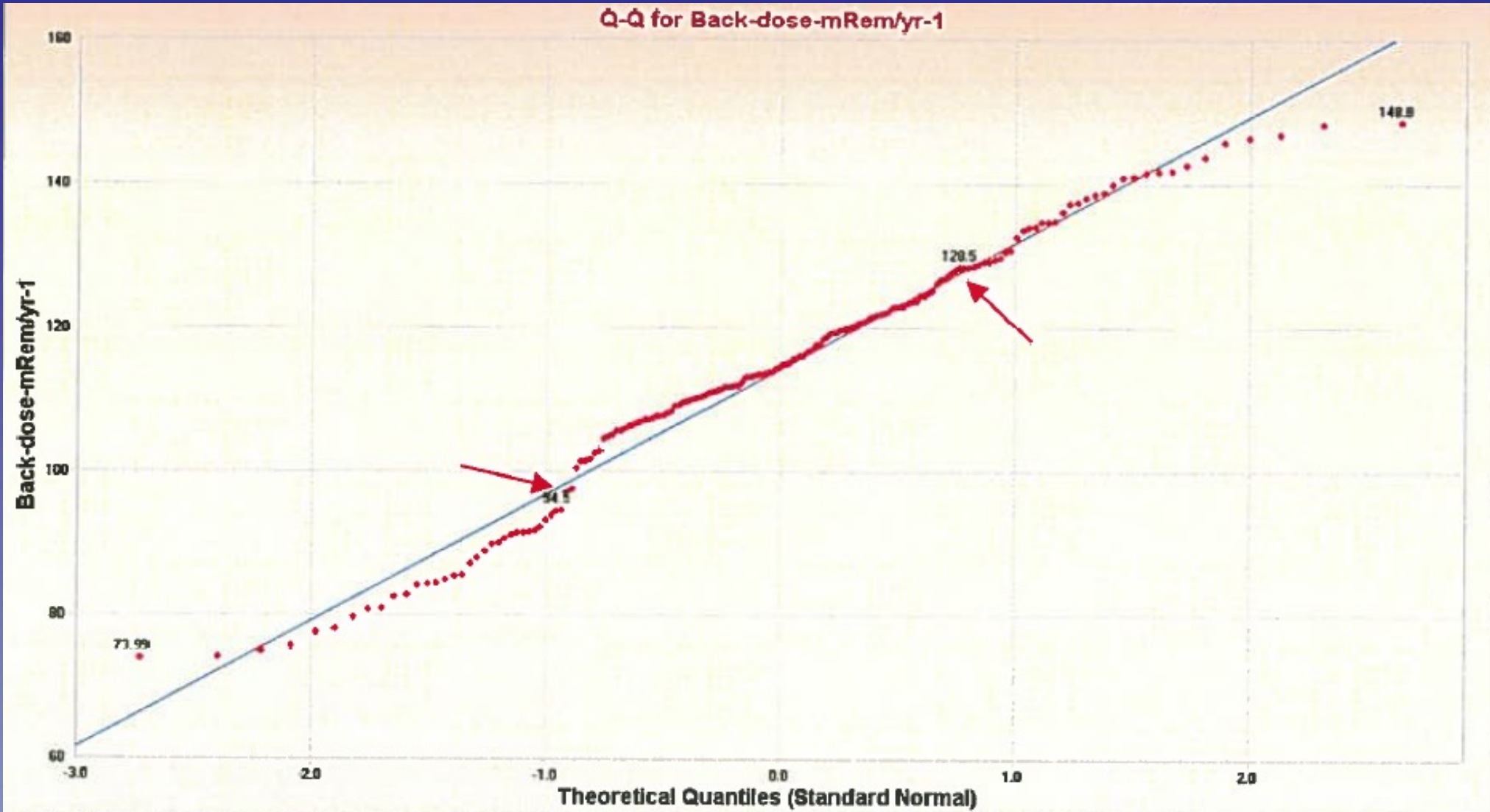


# Goodness of Fit (GOF)

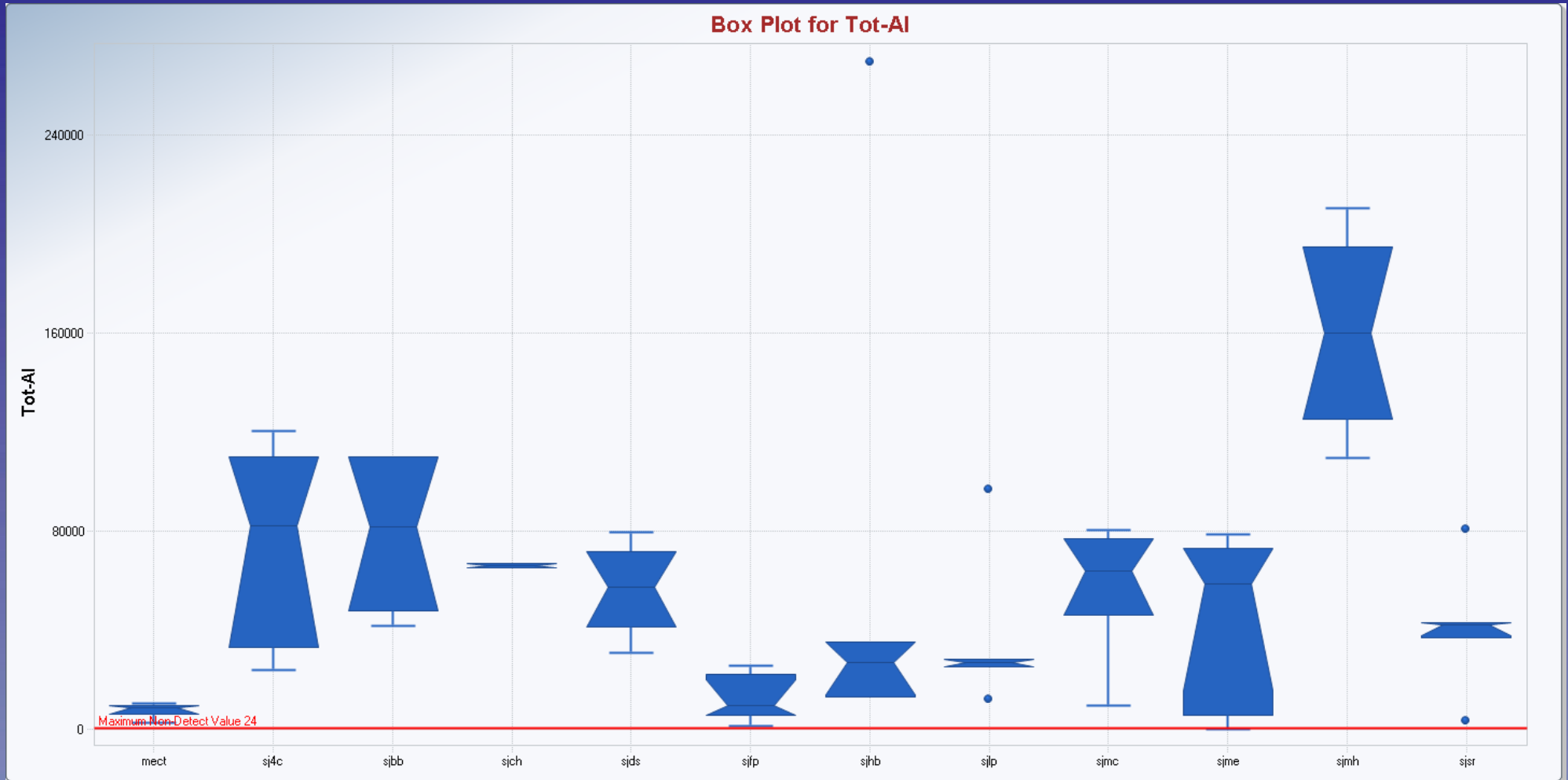
---

- GOF tests the probability that a distribution fits a model; examples:
  - ▶ Shapiro Wilk and Lilliefors for Normal
  - ▶ Shapiro Wilk and Lilliefors for Lognormal
  - ▶ Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S) for Gamma
- Ra-226 data set indicates a 95% probability:
  - ▶ Not Normal
  - ▶ Approximately Lognormal
  - ▶ Gamma distributed

# Multiple Data Populations



# Multiple Data Populations





A stylized logo consisting of a large circle with a smaller circle inside it, positioned at the top. Below the large circle are two leaf-like shapes, one on the left and one on the right, pointing downwards. The word "Outliers" is written in yellow text across the middle of the large circle.

**Outliers**

# Remove Outliers

Ra-226 Background Data Set of Size 20  $\rightarrow$  18

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	<del>19)</del>	<del>22.11</del>
10)	7.38	<del>20)</del>	<del>35.87</del>

Mean = 8.53

↓  
6.26

Median = 7.38

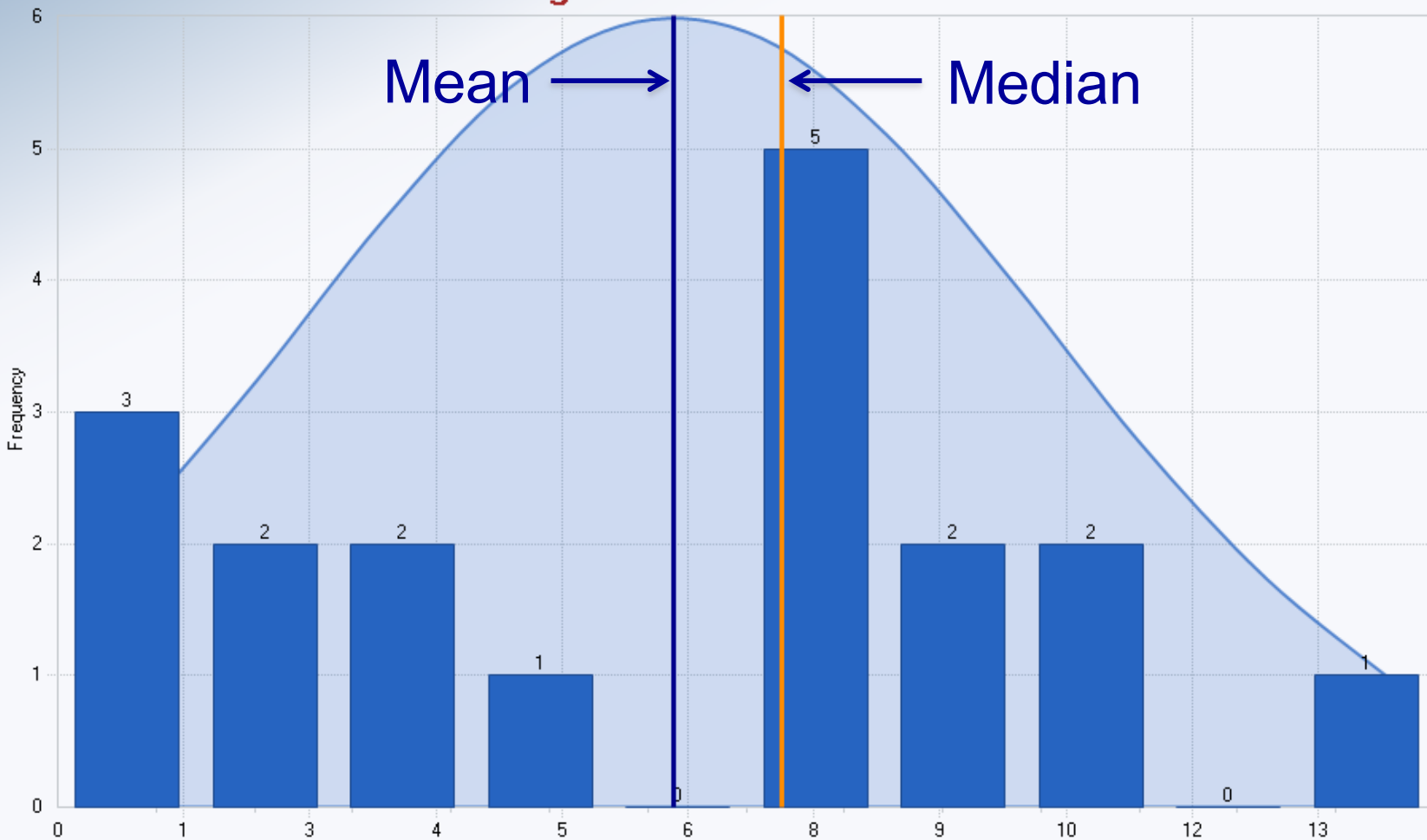
↓  
7.37

Standard  
Deviation = 8.18

↓  
3.82

# Histogram Without Outliers

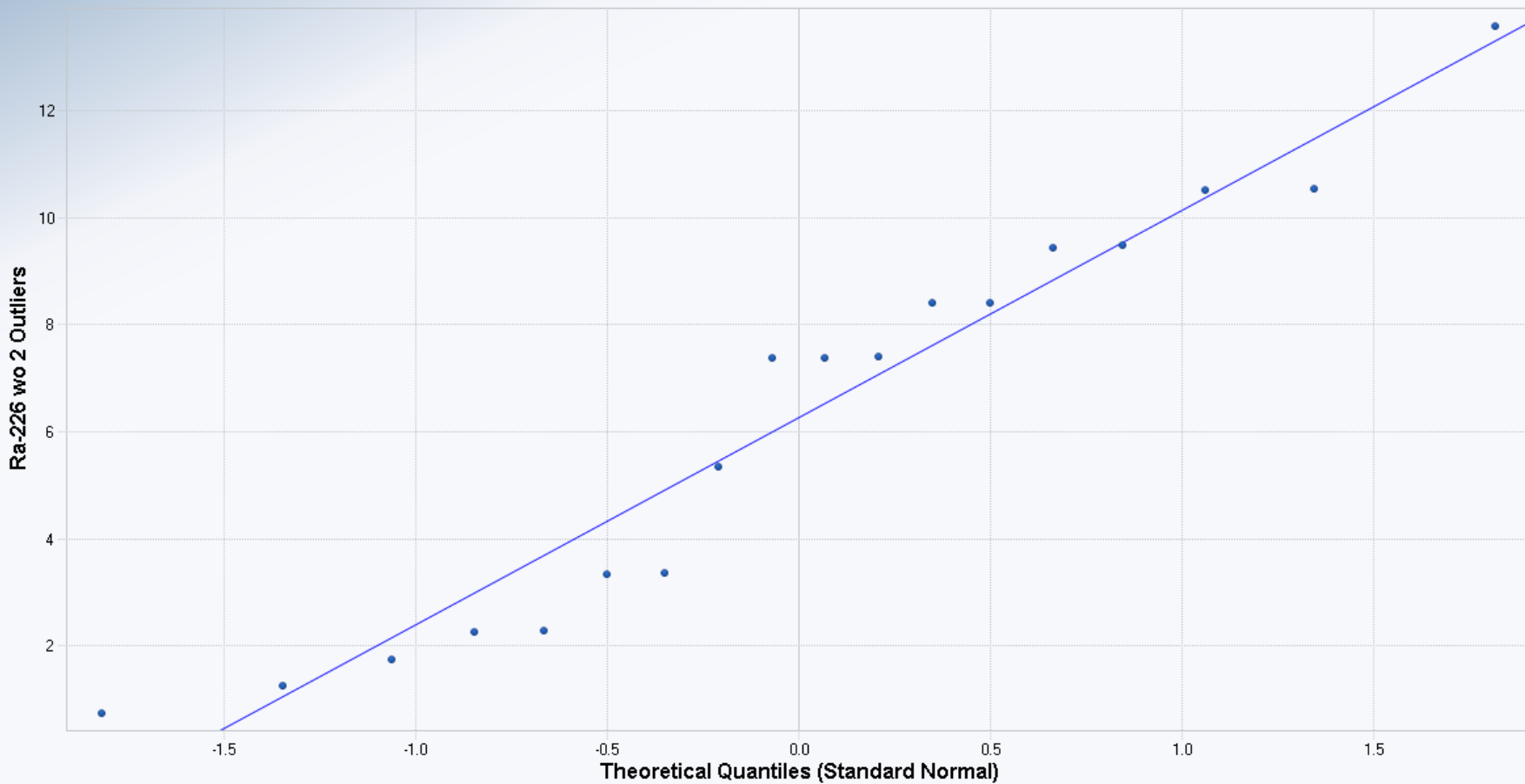
Histogram for Ra-226 wo 2 Outliers



Ra-226 wo 2 Outliers	
Number of Values	18
Minimum	0.74
Maximum	13.56
SD	3.82
Skewness	0.08
Kurtosis	-1.11
<input checked="" type="checkbox"/> Mean	6.26
<input checked="" type="checkbox"/> Median	7.37
<input checked="" type="checkbox"/> Normal Distribution	
<input type="checkbox"/> Less Bins	
<input type="checkbox"/> More Bins	

# Q-Q Plot Without Outliers

Q-Q Plot for Ra-226 wo 2 Outliers

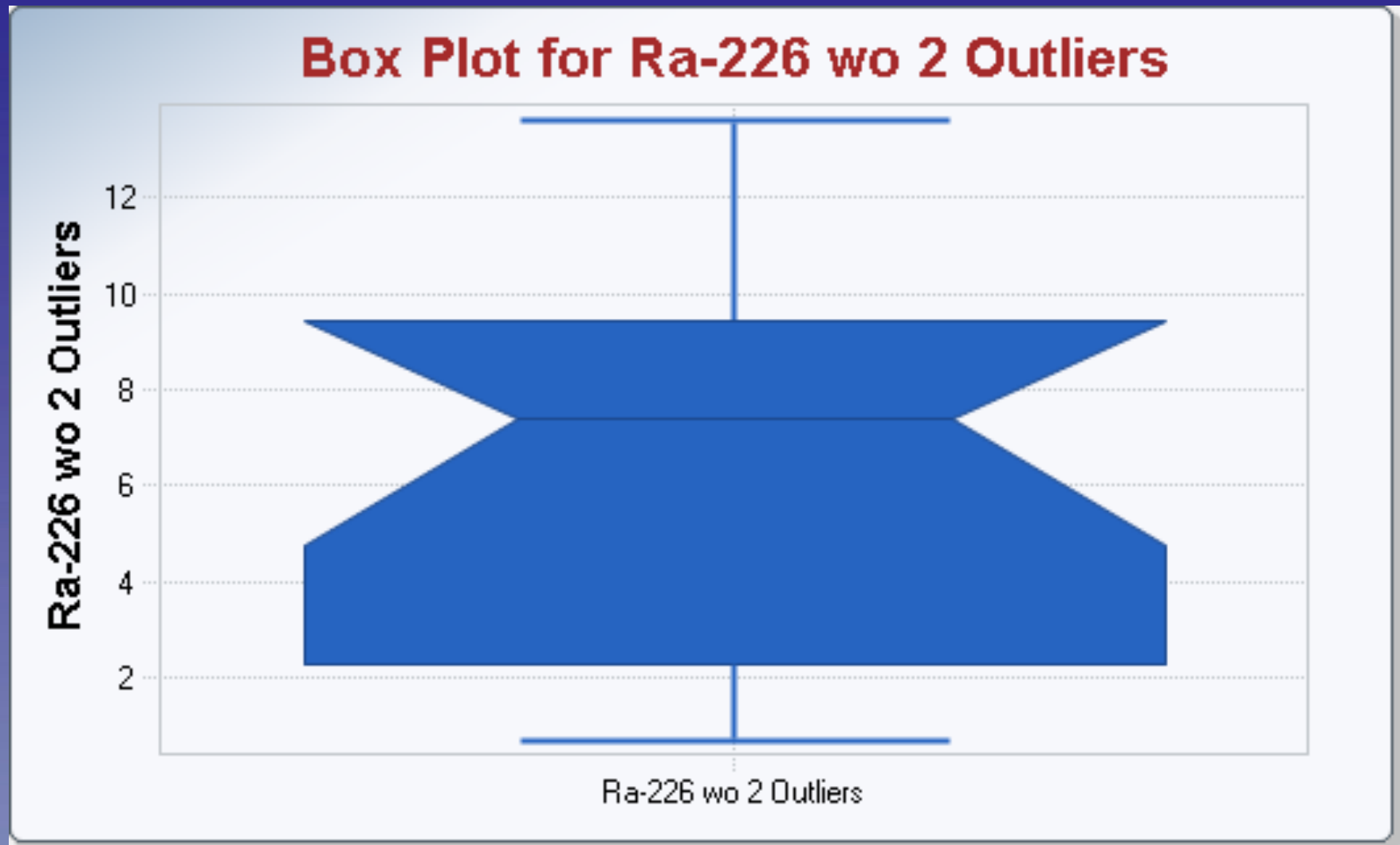


## Ra-226 wo 2 Outliers

N = 18  
Mean = 6.262  
Sd = 3.823  
Slope = 3.874  
Intercept = 6.262  
Correlation, R = 0.973

Best Fit Line

# Box Plot Without Outliers



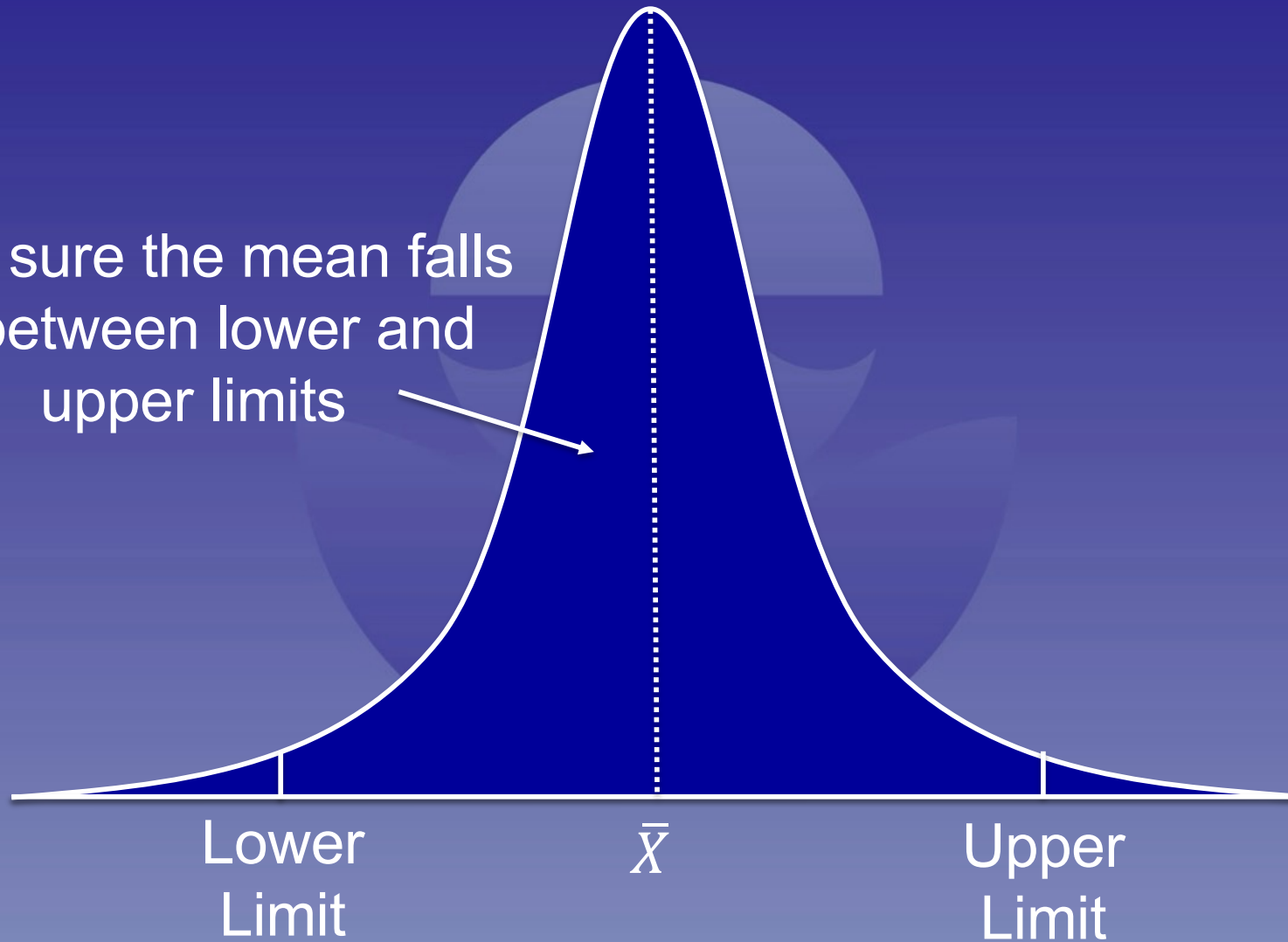


# UCLs and BTVs

# Confidence Interval

---

95% sure the mean falls  
in between lower and  
upper limits



# UCLs and BTVs

---

- Upper Confidence Limit (UCL): Upper limit of a confidence interval for a parameter of interest (typically the mean).
- 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
<del>Lognormal</del>	<del>18.3</del>	<del>13.0</del>	<del>68.6</del>	<del>39.9</del>
Non-Parametric	11.6	7.74	35.9	13.6

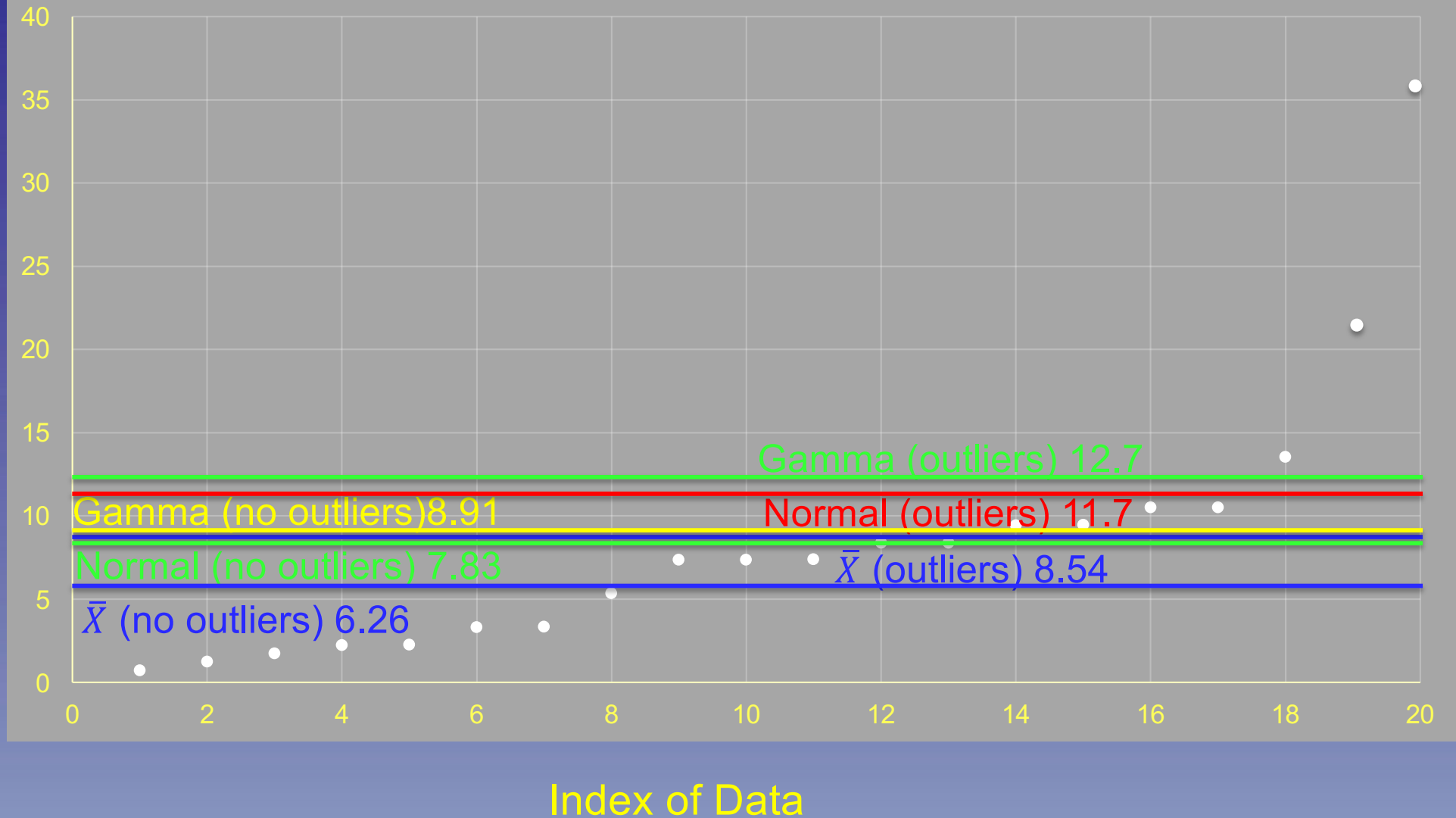
Green = good fit

Yellow = approximate fit

Red = bad fit

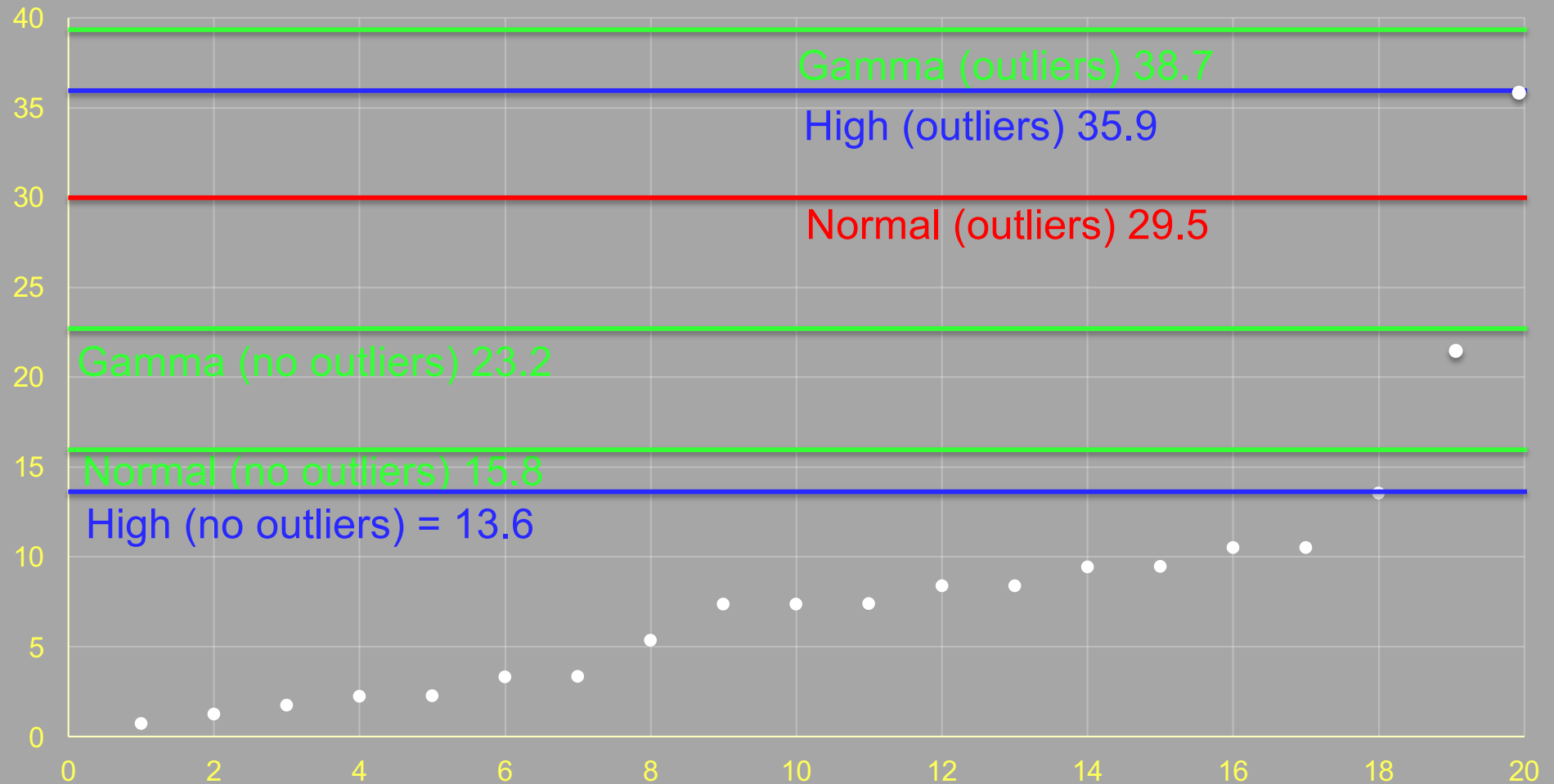
# Graphic Comparison of UCLs

Index Plot of Background Compared to UCLs



# Graphic Comparison of BTVs

Index Plot of Background Compared to BTVs



Index of Data

A stylized logo in a light blue color, centered on the slide. It depicts a person's head and shoulders, with a circular shape representing the head and a smaller circle inside representing the face. Below the head is a shape resembling a flower or a stylized plant with two large, rounded leaves.

# Hypothesis Testing & Decision Errors

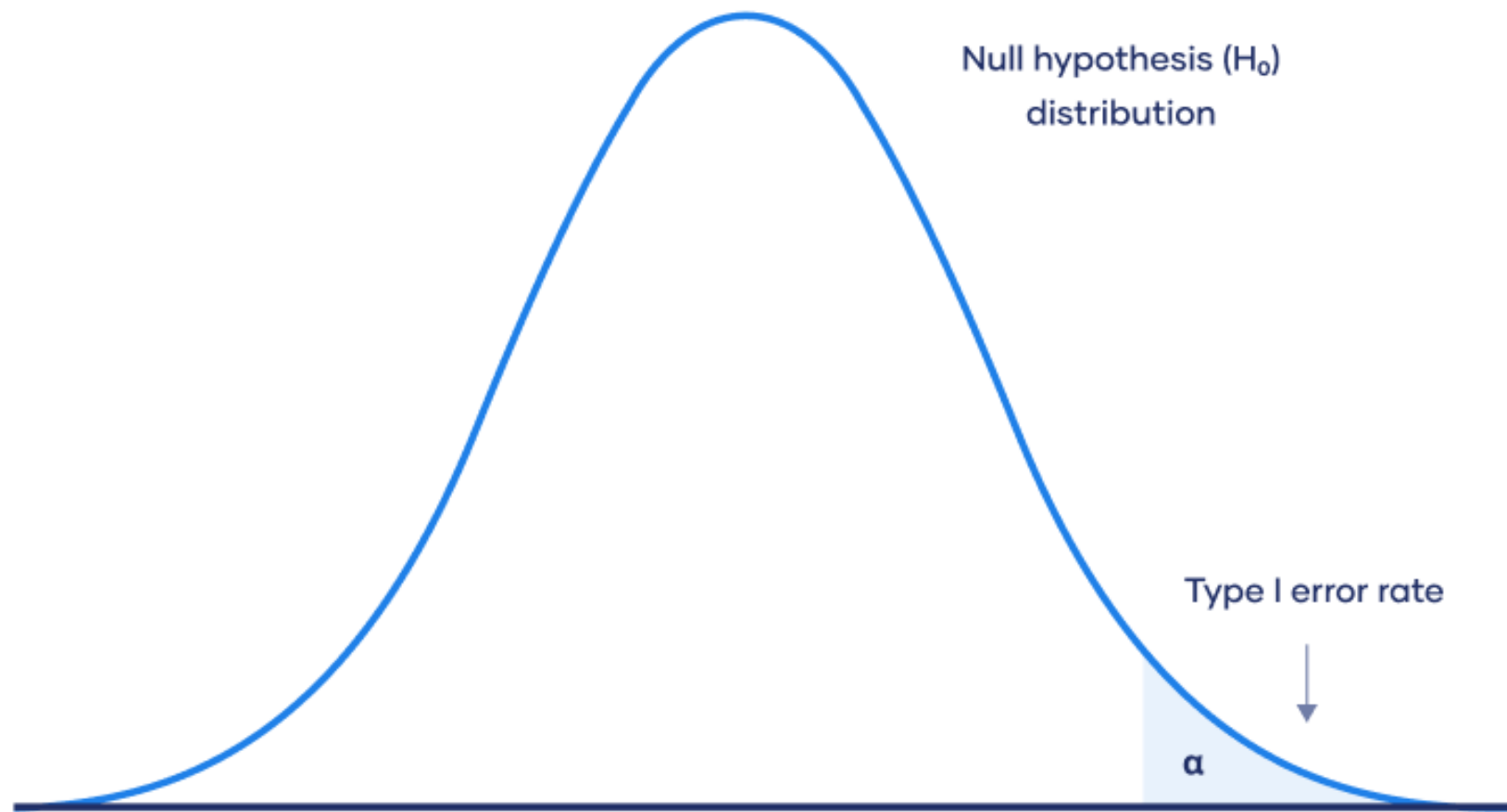
# Hypothesis Testing

---

- Null hypothesis ( $H_0$ ): The site is contaminated.
- Alternative hypothesis ( $H_a$ ): The site is not contaminated.
- Decision errors:
  - ▶ Type I –  $H_0$  is true, but we say false
  - ▶ Type II –  $H_0$  is false, but we say true
- We want to control Type I errors the most.

# Alpha

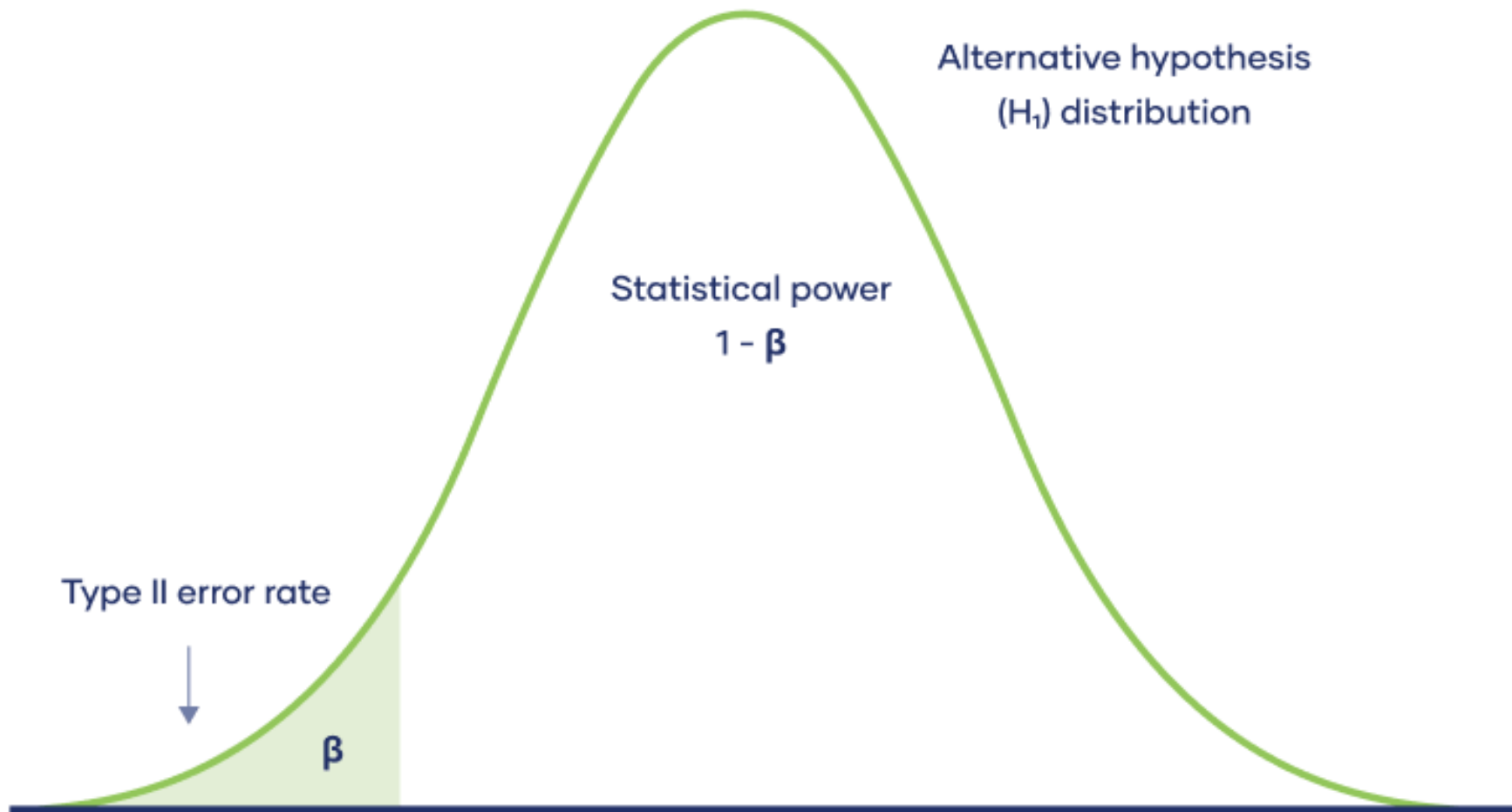
## Probability of making a Type I error



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

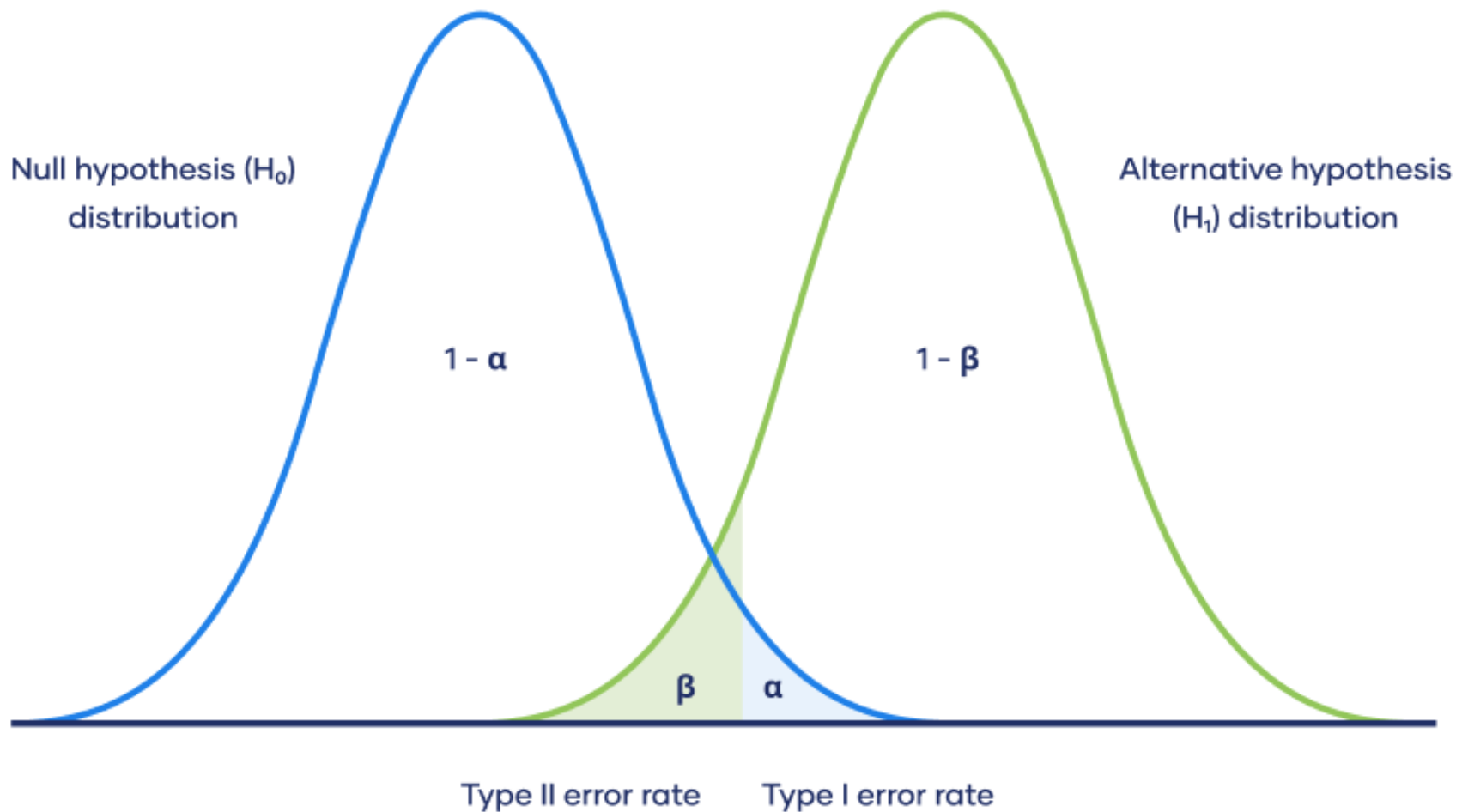
## Probability of making a Type II error



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

## Probability of making Type I and Type II errors





# Decision Errors Can Be Serious!

---

$H_0$  = You are pregnant

Type I Error



Type II Error



# Decision Error Rates

---

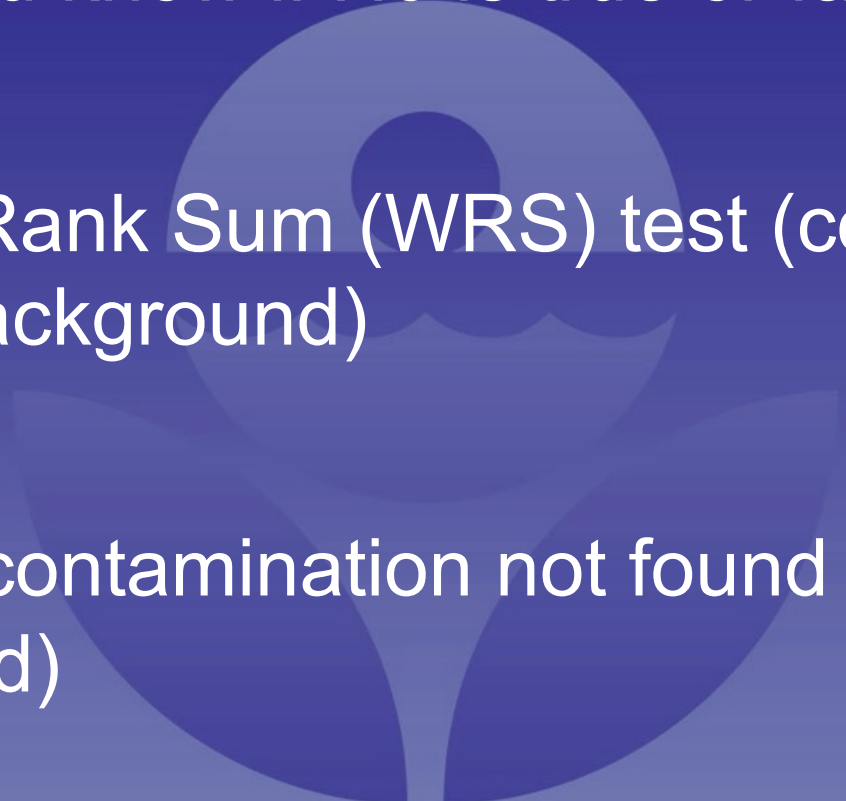
- Alpha is set at 5% (typically)
- Beta is set at 10% (typically)
- How much data you need to collect to determine if  $H_0$  is true or false is dependent on alpha and beta (plus a few other variables)

▶  $N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{3(P_r - 0.5)^2}$  (contamination found in background)

▶  $N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{4(\text{Sign } p - 0.5)^2}$  (contamination not found in background)

# Statistical Testing

---

- How do you know if  $H_0$  is true or false?
  - Wilcoxon Rank Sum (WRS) test (contamination found in background)
  - Sign test (contamination not found in background)
- 

# 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%?                      75%?  
2,000%?                      23%?                      64%?  
350%?



**ProUCL**

# ProUCL Menu

ProUCL 5.1 - [Example\_Ra-226\_Background\_Data\_Set.xls]

File Edit Stats/Sample Sizes Graphs Statistical Tests Upper Limits/BTVs UCLs/EPCs Windows Help

Navigation Panel

Name  
Example\_Ra-226\_Backgr

	0	1	2	3	4	5
	Ra-226	Ra-226 wo 1 Outlier	Ra-226 wo 2 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						

# Graphs

The screenshot shows the ProUCL 5.1 software interface. The main window displays a data table with 22 rows and 5 columns. The first column contains row numbers (1-22). The second column contains data values, and the third column contains values that appear to be the same as the second column. The fourth and fifth columns contain values 2, 3, 4, and 5 respectively. A 'Graphs' menu is open, showing options: Box Plot, Multiple Box Plots, Histogram, Multiple Histograms, Q-Q Plots (highlighted), and Multiple Q-Q Plots. A 'Navigation Panel' on the left shows the file name 'Example\_Ra-226\_Backgr'. The title bar reads 'ProUCL 5.1 - [Example\_Ra-226\_Background\_Data\_Set.xls]'. The menu bar includes 'File', 'Edit', 'Stats/Sample Sizes', 'Graphs', 'Statistical Tests', 'Upper Limits/BTVs', 'UCLs/EPCs', 'Windows', and 'Help'.

	2	3	4	5
Ra-226 wo 2 Outliers				
1				
2				
3				
4				
5				
6	3.33	3.33		
7	3.36	3.36		
8	5.35	5.35		
9	7.36	7.36		
10	7.38	7.38		
11	7.39	7.39		
12	8.4	8.4		
13	8.4	8.4		
14	9.43	9.43		
15	9.47	9.47		
16	10.51	10.51		
17	10.52	10.52		
18	13.56	13.56		
19	22.11	22.11		
20	35.87			
21				
22				

# Option to Select Multiple Data Sets

ProUCL 5.1 - [Example\_Ra-226\_Background\_Data\_Set.xls]

File Edit Stats/Sample Sizes Graphs Statistical Tests Upper Limits/BTVs UCLs/EPCs Windows Help

Navigation Panel

	0	1	2	3	4	5	6	7	8	9	10	11
Name	Ra-226	Ra-226 wo 1 Outlier	Ra-226 wo 2 Outliers									
Example_Ra-226_Backgr	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											

Select Variables

Available Variables

Name	ID
Ra-226	0
Ra-226 wo 1 Ou...	1
Ra-226 wo 2 Ou...	2

Selected Variables

Name	ID
------	----

Select Group Column (Optional)

Options OK Cancel



# Statistical Tests

The screenshot shows the ProUCL 5.1 software interface. The 'Statistical Tests' menu is open, and 'Goodness-of-Fit Tests' is selected. A sub-menu is visible, listing 'Normal', 'Gamma', 'Lognormal', and 'G.O.F. Statistics'. The background shows a data table with 20 rows and 4 columns.

			4	5	6
1					
2					
3					
4	2.25	2.25			
5	2.28	2.28			
6	3.33	3.33			
7	3.36	3.36			
8	5.35	5.35			
9	7.36	7.36			
10	7.38	7.38			
11	7.39	7.39			
12	8.4	8.4			
13	8.4	8.4			
14	9.43	9.43			
15	9.47	9.47			
16	10.51	10.51			
17	10.52	10.52			
18	13.56	13.56			
19	22.11	22.11			
20	35.87				
21					
22					
23					
24					

# “Options” Can Be Changed

The screenshot displays the ProUCL 5.1 software interface. The main window shows a data table with columns labeled 0 through 12. The data is organized into three groups: Ra-226 (column 0), Ra-226 wo 1 Outlier (column 1), and Ra-226 wo 2 Outliers (column 2). The data values are as follows:

	0	1	2	3	4	5	6	7	8	9	10	11	12
	Ra-226	Ra-226 wo 1 Outlier	Ra-226 wo 2 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													

Two dialog boxes are overlaid on the data table:

- Select Variables**: A dialog box with two columns: "Available Variables" and "Selected Variables". The "Available Variables" column lists "Ra-226" (ID 0), "Ra-226 wo 1 Ou..." (ID 1), and "Ra-226 wo 2 Ou..." (ID 2). The "Selected Variables" column is currently empty.
- GOF\_ConfLevelForm**: A dialog box titled "Select Confidence Coefficient" with three radio buttons: "99%", "95%" (selected), and "90%". It includes "OK" and "Cancel" buttons.

# BTV Menu

The screenshot shows the ProUCL 5.1 software interface. The menu bar includes File, Edit, Stats/Sample Sizes, Graphs, Statistical Tests, Upper Limits/BTVs, UCLs/EPCs, Windows, and Help. The 'Upper Limits/BTVs' menu is open, displaying options: Normal, Gamma, Lognormal, Non-Parametric, and All (highlighted). The main window displays a data table with columns for '0', 'Ra-226', and 'Ra-2'. The table contains 22 rows of data. A 'Navigation Panel' on the left shows the file name 'Example\_Ra-226\_Backgro'.

	0	Ra-226	Ra-2
1		0.74	
2		1.24	1.24
3		1.75	1.75
4		2.25	2.25
5		2.28	2.28
6		3.33	3.33
7		3.36	3.36
8		5.35	5.35
9		7.36	7.36
10		7.38	7.38
11		7.39	7.39
12		8.4	8.4
13		8.4	8.4
14		9.43	9.43
15		9.47	9.47
16		10.51	10.51
17		10.52	10.52
18		13.56	13.56
19		22.11	22.11
20		35.87	
21			
22			

# UCL Menu

The screenshot shows the ProUCL 5.1 software interface. The main window displays a data table with columns for '0 Ra-226', '1 Ra-226 wo 1 Outlier', and 'Ra-226'. The 'UCLs/EPCs' menu is open, showing options: Normal, Gamma, Lognormal, Non-Parametric, and All. The 'All' option is selected. The data table contains 22 rows of data.

	0 Ra-226	1 Ra-226 wo 1 Outlier	Ra-226
1	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			



**Questions?**