ProUCL Utilization 2020
ProUCL A to Z

Presenters:

Travis Linscome-Hatfield,

Anita Singh

Polona Carson

NEPTUNE
AND COMPANY
Improving the quality of
environmental decision making.

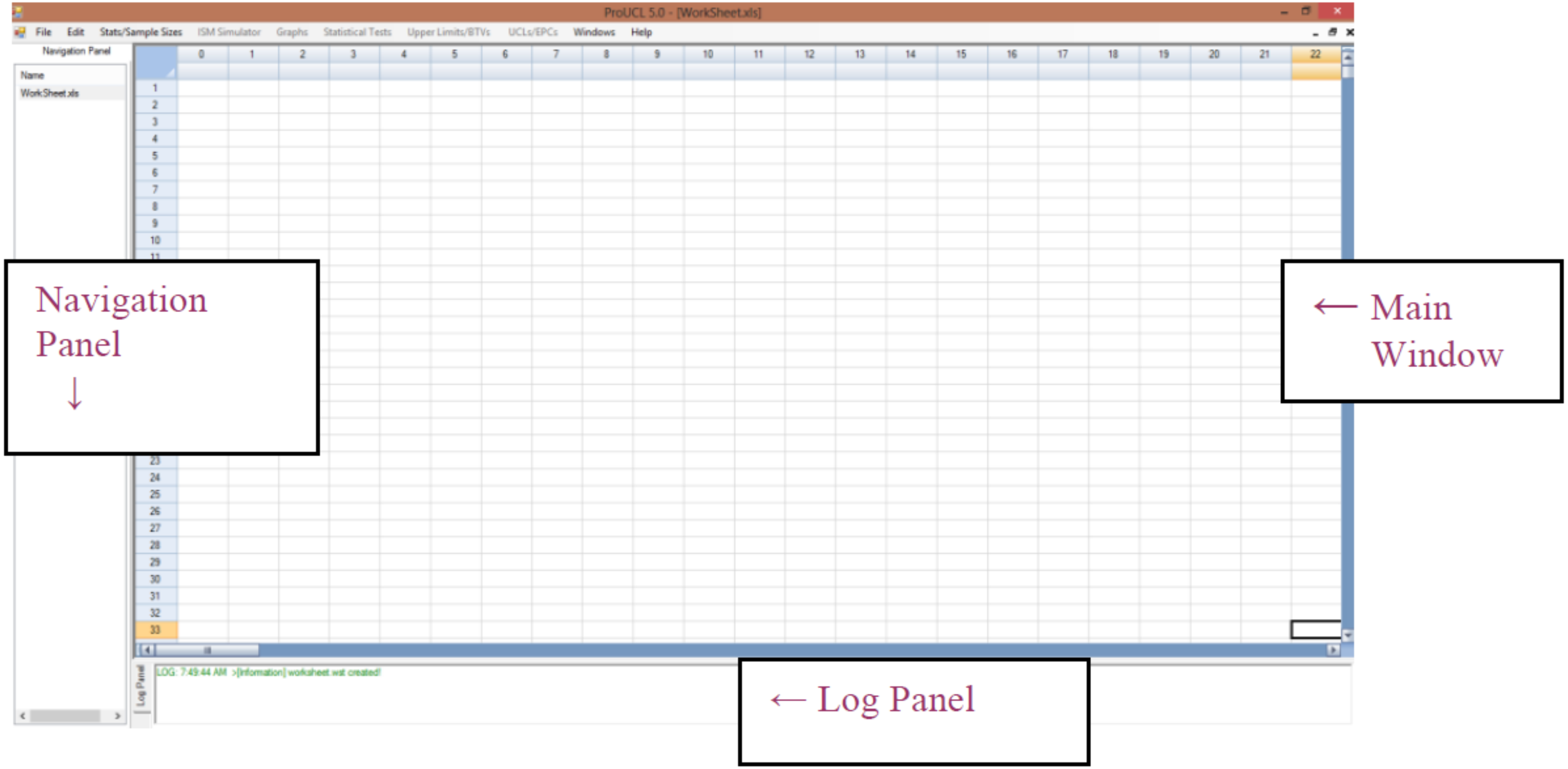
Learning objectives

- Objectives
  - Get familiar with ProUCL and some commonly used data analysis features
- Today we will discuss:
  - Starting ProUCL
  - Preparing data for analysis and loading in ProUCL
  - Basics of dealing with missing values and NDs
  - Exploratory Data Analysis
  - Hypothesis testing

ProUCL Software

- Statistical software for environmental data analysis
- User Guide
  - Provides instructions on how to use ProUCL
- Technical Guide
  - Provides detailed background on statistical methods

Navigating ProUCL



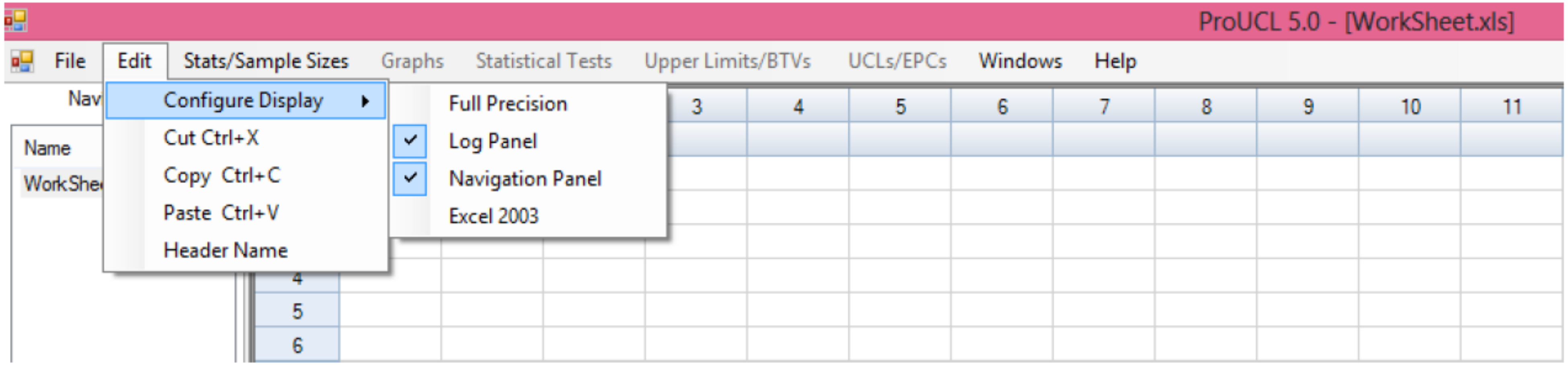
```
LOG: 7:49:44 AM >[Information] worksheet.wst created!
```

Navigation Panel ↓

← Main Window

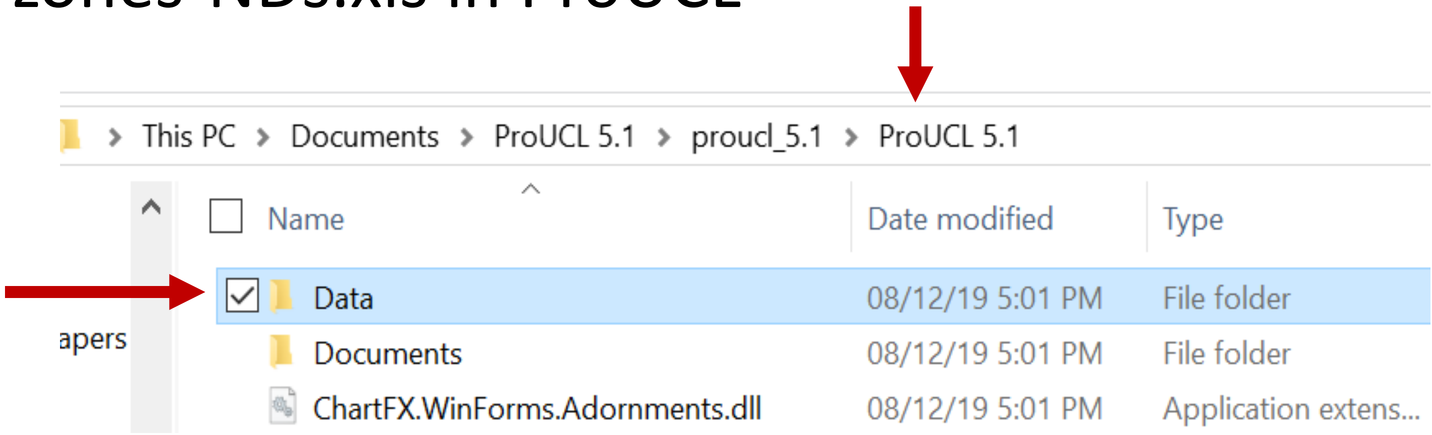
← Log Panel

Turning panels on / off



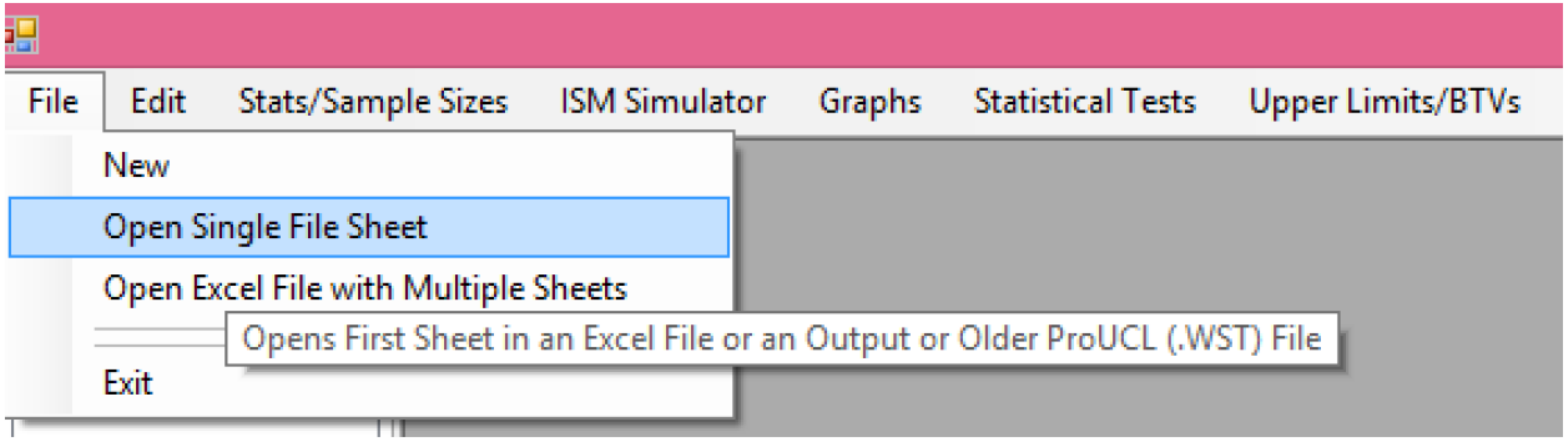
Starting ProUCL and Loading the data

- Zn-Cu-two-zones-NDs.xls in ProUCL



› This PC › Documents › ProUCL 5.1 › proucl_5.1 › ProUCL 5.1

□ Name	Date modified	Type
☑ Data	08/12/19 5:01 PM	File folder
Documents	08/12/19 5:01 PM	File folder
ChartFX.WinForms.Adornments.dll	08/12/19 5:01 PM	Application extens...



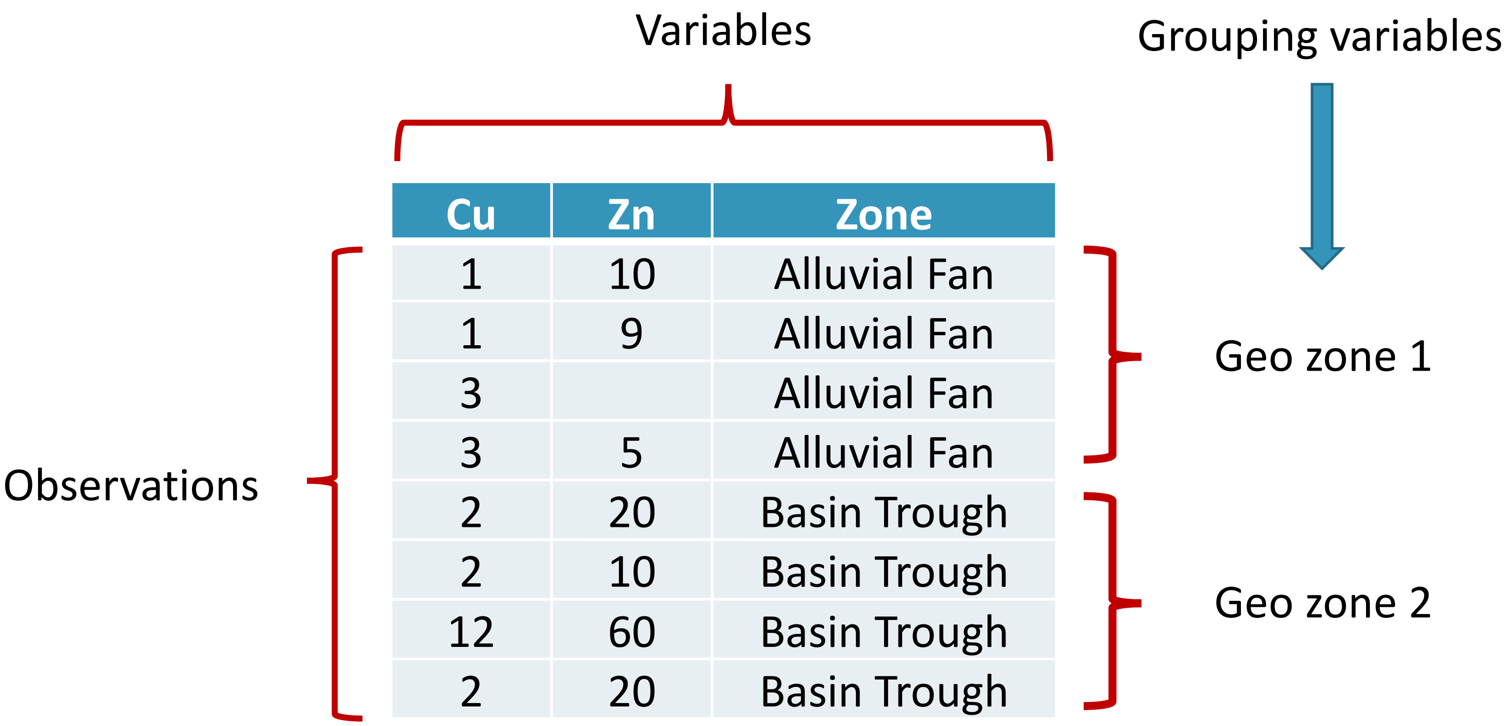
Opens First Sheet in an Excel File or an Output or Older ProUCL (.WST) File

Data set

- Zn-Cu-two-zones-NDs.xls available in ProUCL 5.1 Data folder
- Copper and zinc concentrations (mg/L) in shallow ground water from two geological zones (Alluvial Fan and Basin-Trough) in the San Joaquin Valley, CA.
- Multiple detection limits for both the copper and zinc data
  - at 1, 2, 5, 10 and 20 ug/L
- Original source:
  - **Millard, S.P. and Deverel, S.J. (1988).** Nonparametric statistical methods for comparing two sites based on data with multiple non-detect limits. *Water Resources Research 24: doi: 10.1029/88WR03412. issn: 0043-1397*

How to organize data?

- Columns → variables
- Rows → observations
- Grouping variable
  - Count denotes iris species
  - Equal counts
- Data formats
  - .xlsx (Excel)
  - .xls (Excel)
  - .wst (Worksheet)
  - .ost (Output)



Variables

Grouping variables

Observations

Cu	Zn	Zone
1	10	Alluvial Fan
1	9	Alluvial Fan
3		Alluvial Fan
3	5	Alluvial Fan
2	20	Basin Trough
2	10	Basin Trough
12	60	Basin Trough
2	20	Basin Trough

Geo zone 1

Geo zone 2

Nondetects

- Nondetect (ND) values
  - Censored data values
  - Concentrations or measurements that are less than the analytical/instrument method detection limit or reporting limit.
- How to designate nondetect values?
  - Add new variable for each variable with nondetects
  - Column name: d_ + variable name (Cu → D_Cu)
  - No missing values in d- column!!

1 = detect, 0 = nondetect

Cu	Zn	Zone	D_Cu	D_Zn
1	10	Alluvial Fan	0	0
1	9	Alluvial Fan	0	1
3		Alluvial Fan	1	
3	5	Alluvial Fan	1	1

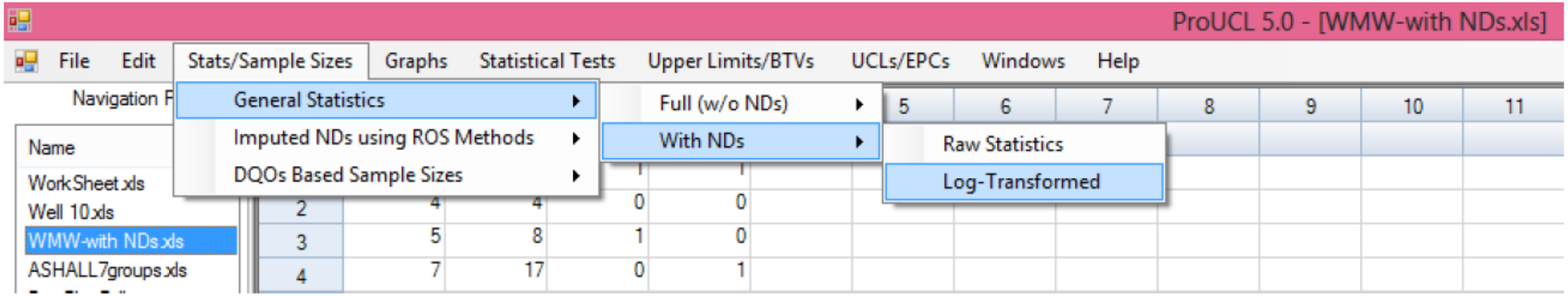
Cu	Zn	Zone	D_Cu	D_Zn
1	10	Alluvial Fan	0	0
	9	Alluvial Fan	0	1
3	no data	Alluvial Fan	1	1e31
3	5	Alluvial Fan	1	1

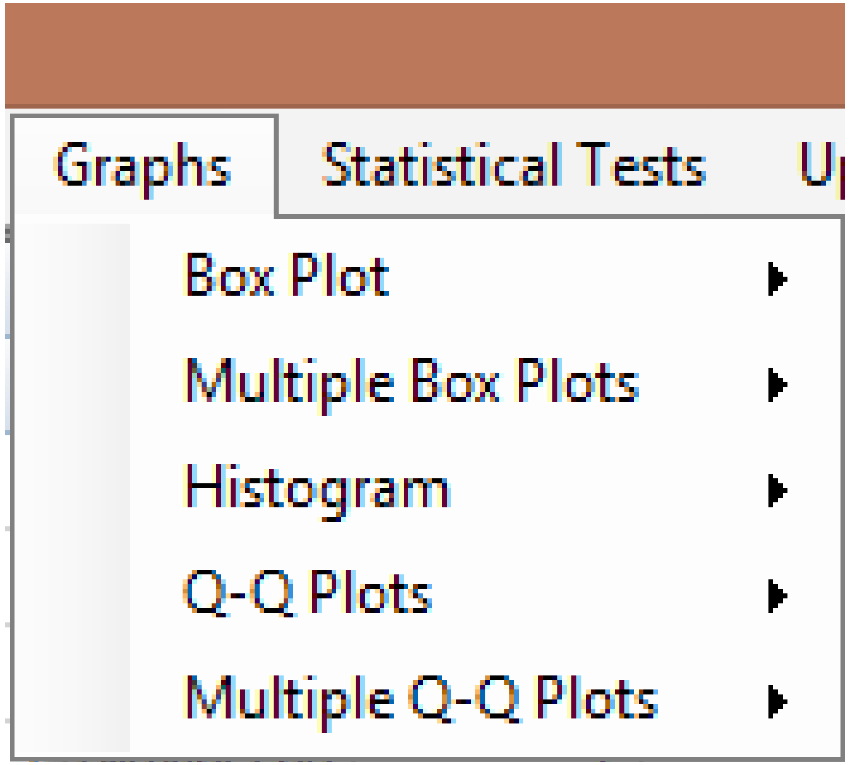
Missing Data

- Blanks
- Alphanumeric strings
- Very large values (1e31)

Exploratory Data Analysis (EDA)

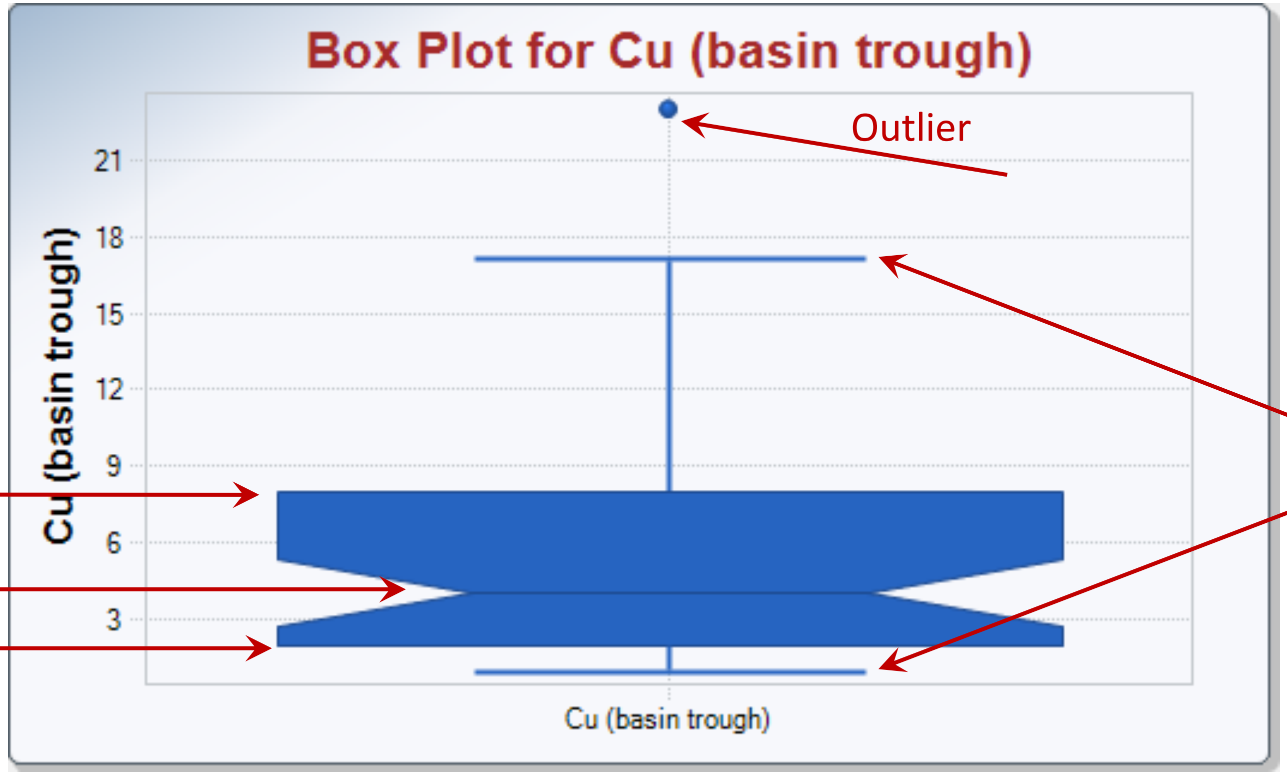
- Summary statistics - User Guide Chapter 4





Exploratory Data Analysis (EDA)-I

- Graphical presentations of data
  - User Guide Chapter 6



Outlier

Fences

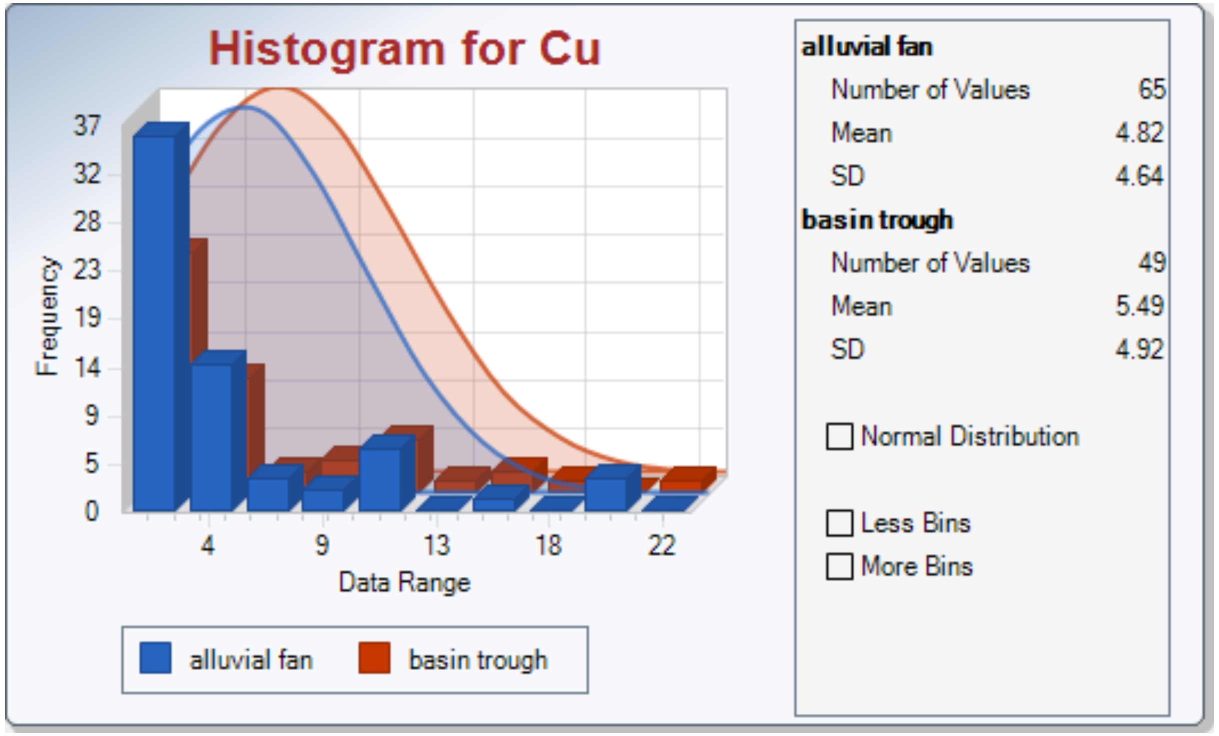
Q3

Q2 / median

Q1

Box Plot

- Quick 5-point summary:
  - Lowest / highest value
  - Median (Q2)
  - Degree of dispersion
  - Degree of skewness
  - Unusual data



alluvial fan	
Number of Values	65
Mean	4.82
SD	4.64
basin trough	
Number of Values	49
Mean	5.49
SD	4.92

☐ Normal Distribution

☐ Less Bins

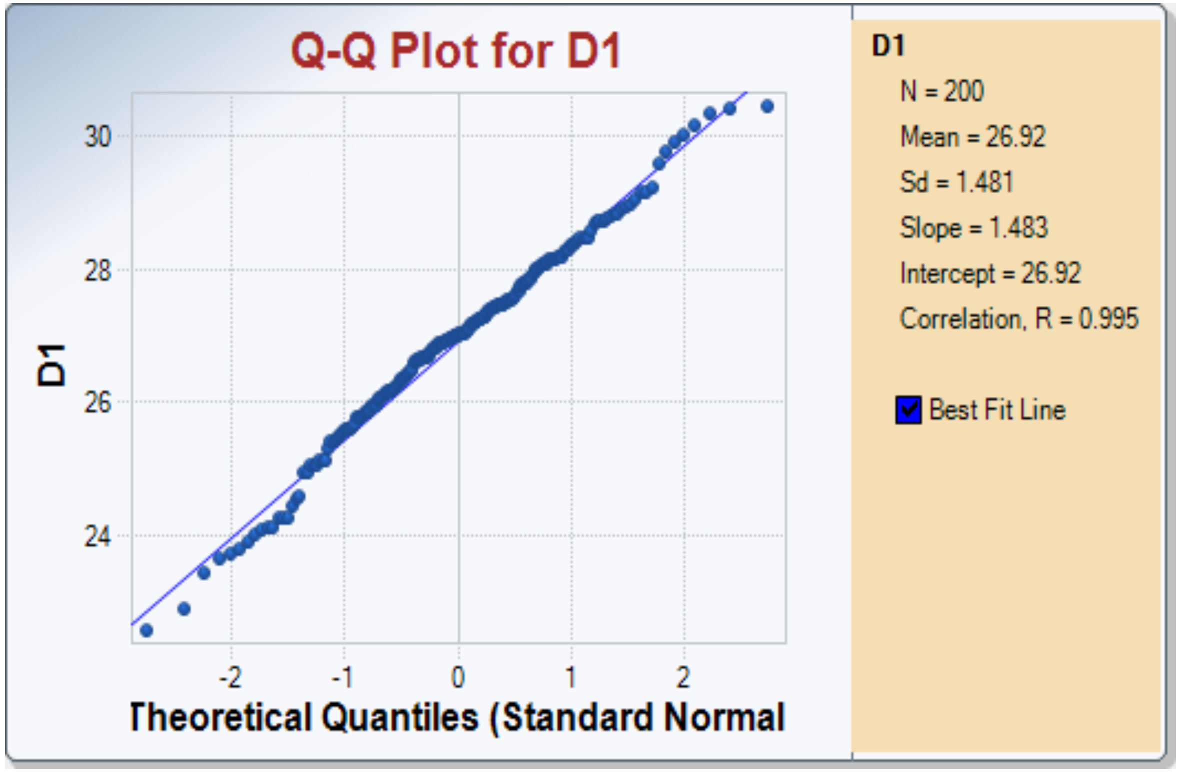
☐ More Bins

Histogram – Cu

- Shape
- Center (location) of the data
- Spread of the data
- Skewness

Q-Q plot

Normally distributed



D1

N = 200

Mean = 26.92

Sd = 1.481

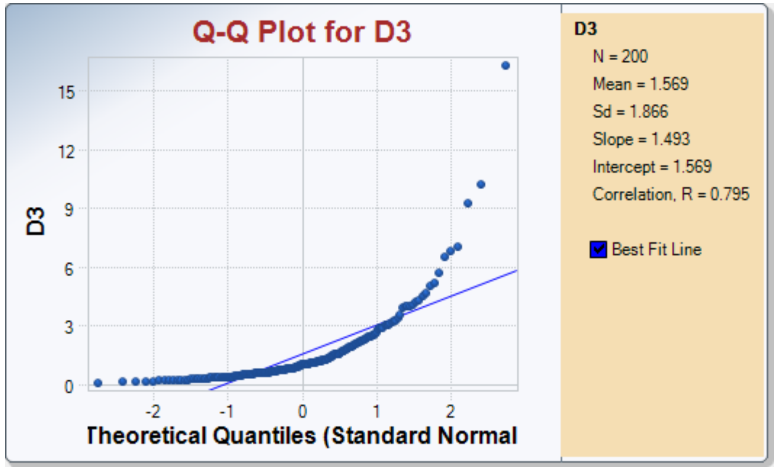
Slope = 1.483

Intercept = 26.92

Correlation, R = 0.995

☑ Best Fit Line

Skewed distribution



D3

N = 200

Mean = 1.569

Sd = 1.866

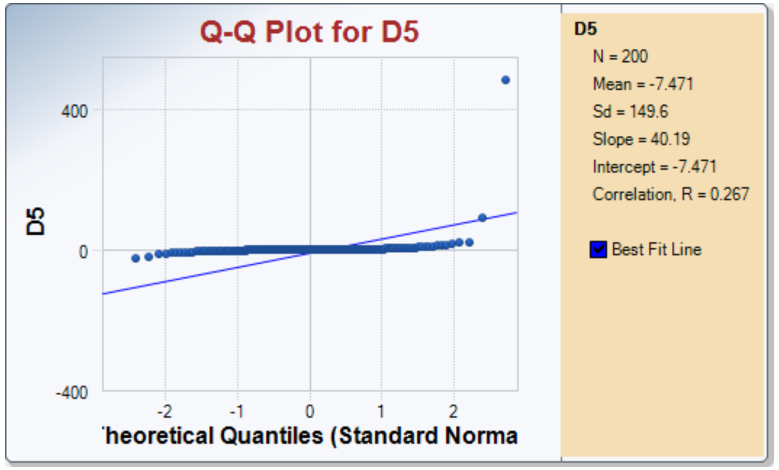
Slope = 1.493

Intercept = 1.569

Correlation, R = 0.795

☑ Best Fit Line

Distribution with heavy tails



D5

N = 200

Mean = -7.471

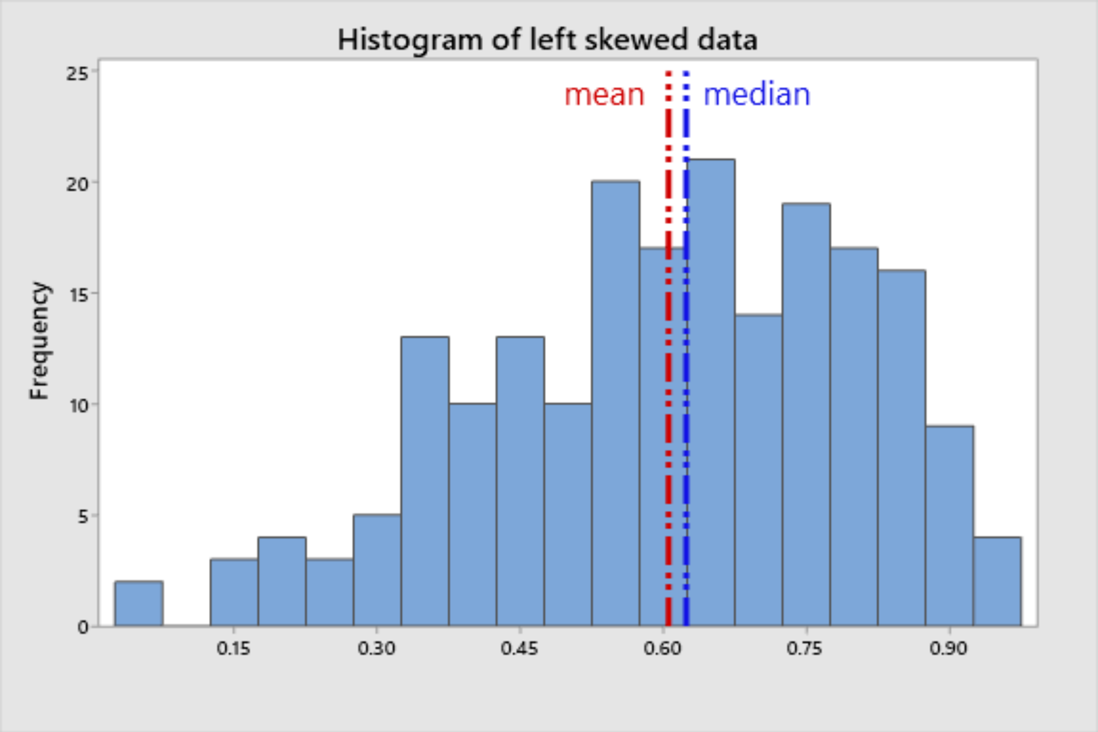
Sd = 149.6

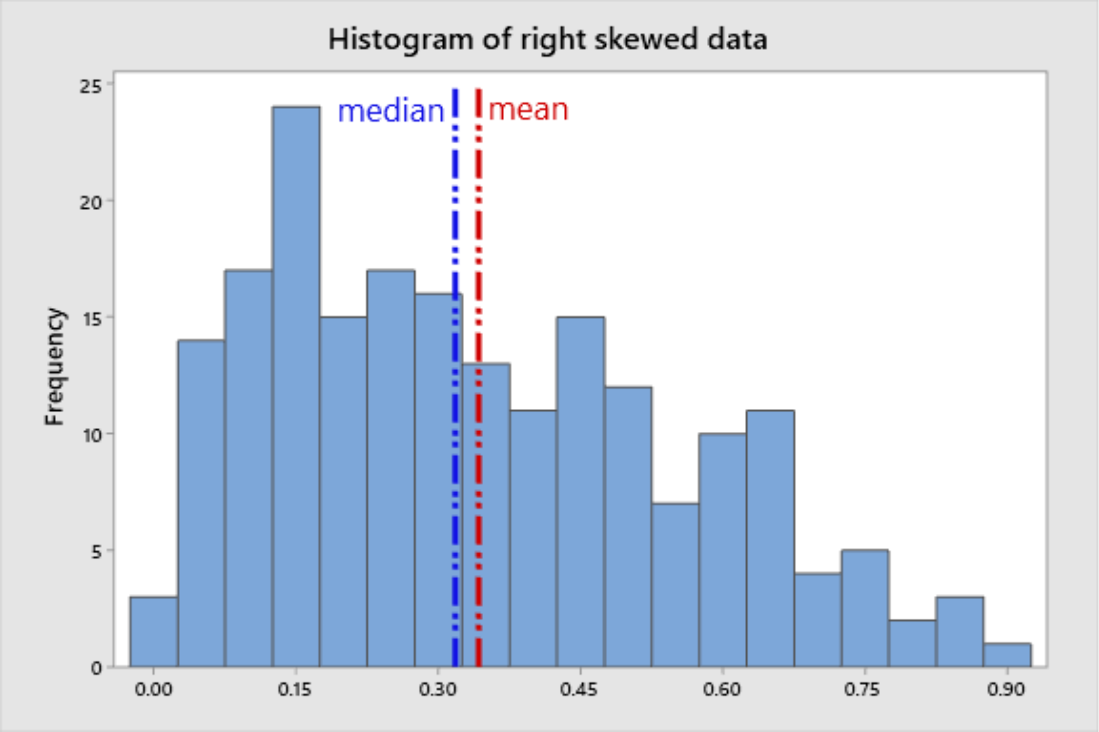
Slope = 40.19

Intercept = -7.471

Correlation, R = 0.267

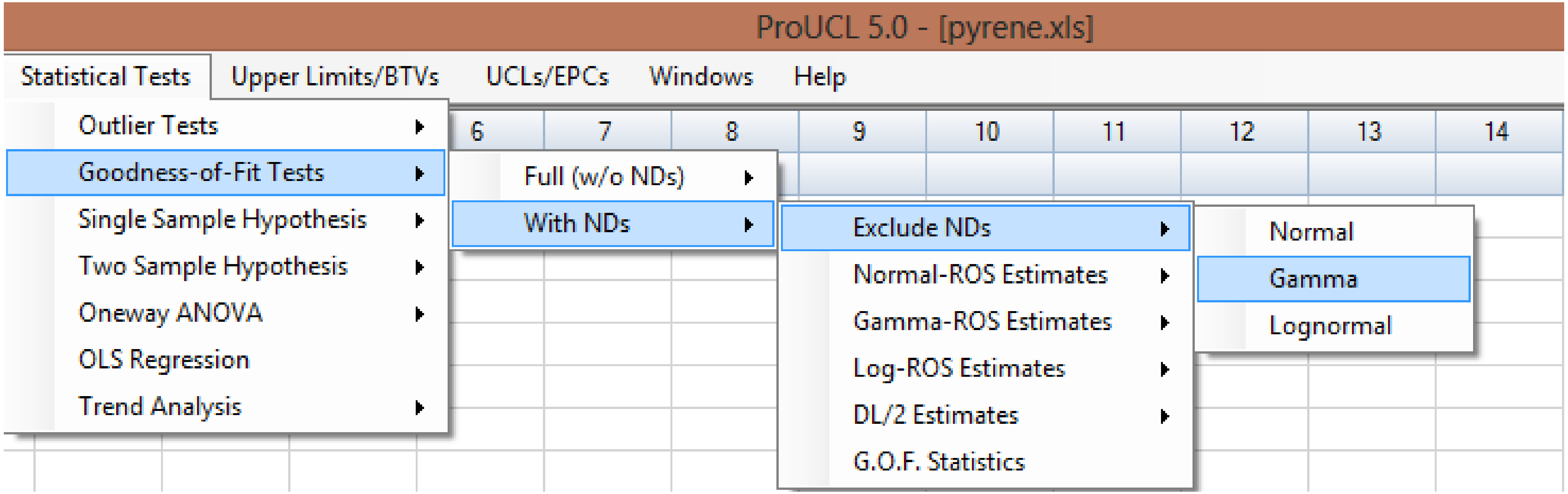
☑ Best Fit Line





Evaluate distribution of the data

- General Statistics Table:
  - Compare Mean & 50% percentile (Median) in General stat table
  - Box plot
  - QQ-plot
  - Goodness of fit test



Goodnes of Fit Test

UG Chapter 8

- Use G.O.F Statistics
- Generates a detailed output
- Helps determine distribution of data set

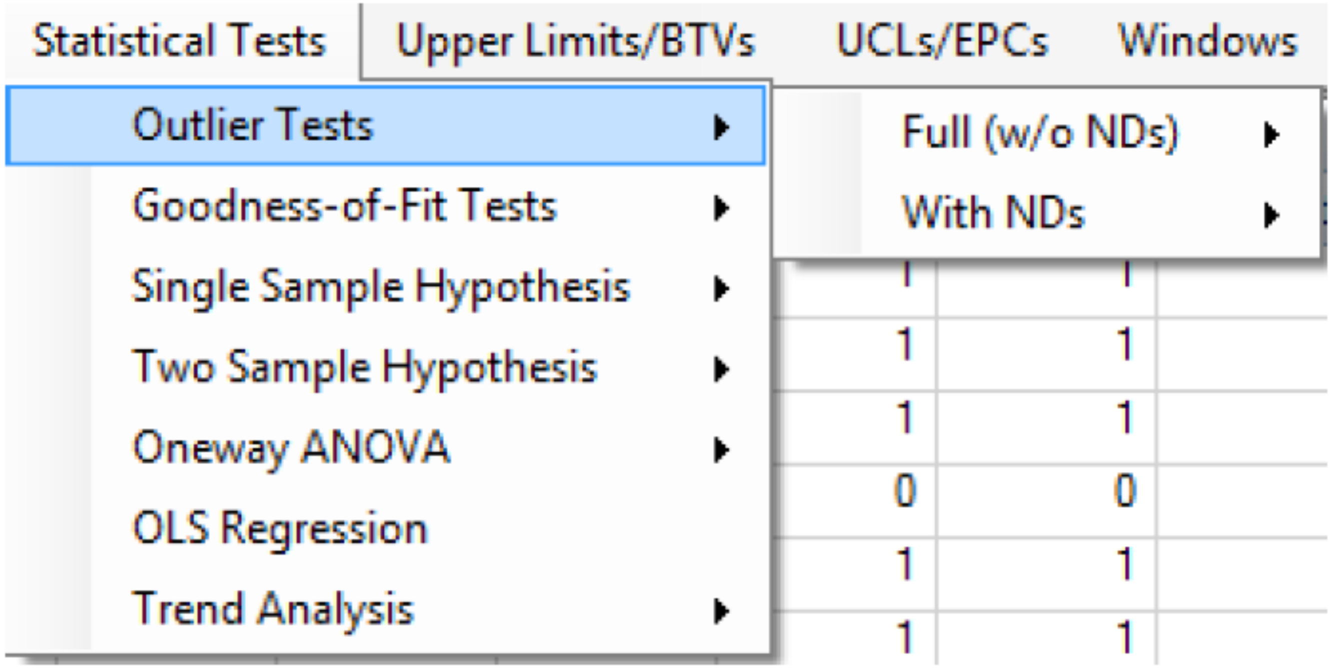
Outliers

- Extremely large or small values relative to the rest of the data
- Suspected to misrepresent the population from which they were collected
- May result from errors:
  - Transcription errors
  - Data-coding errors
  - Laboratory measurement errors
- May indicate more variability than expected
  - Extreme population values
  - On-site hot spots
  - Multiple soil types in background area
- Outliers can distort most decision statistics
  - mean, UCL, UPL, test statistics, …
- “Not removing true outliers or removing false outliers both lead to distorted estimates of population parameters” (QA/G-9S)

Outliers – 5 steps to treat extreme values

1. Identify extreme values that may be potential outliers;
2. Apply statistical test;
3. Scientifically review statistical outliers and decide on their disposition;
4. Conduct data analyses with and without statistical outliers; and
5. Document the entire process.

Reference: EPA guidance QA/G-9S Data Quality Assessment: Statistical Methods for Practitioners



Outlier test –

UG Chapter 7

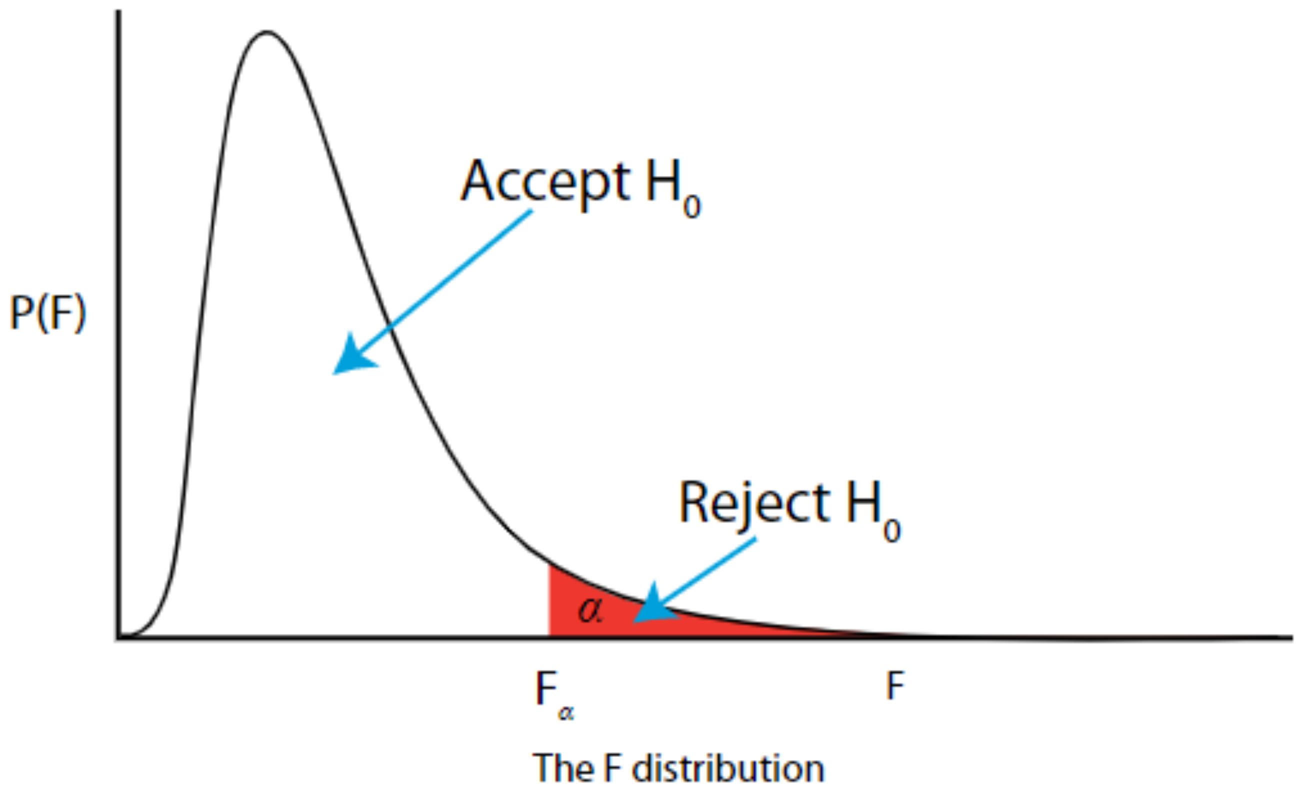
- Dixon and Rosner tests in ProUCL
  - Both require assumption of normality of the data set without outliers
- How to deal with NDs?
  - Exclude NDs
  - Replace NDs b y DL/2 values

Hypothesis testing

- User Guide Chapter 9
- Parametric and non-parametric test are available in ProUCL

- Single-sample hypothesis test
  - To compare site data with pre-specified cleanup standard (Cs) and compliance limit (CL)
- Two-sample hypothesis testing
  - To compare two populations ie: background vs area of concern (AOC)

Steps in hypothesis testing

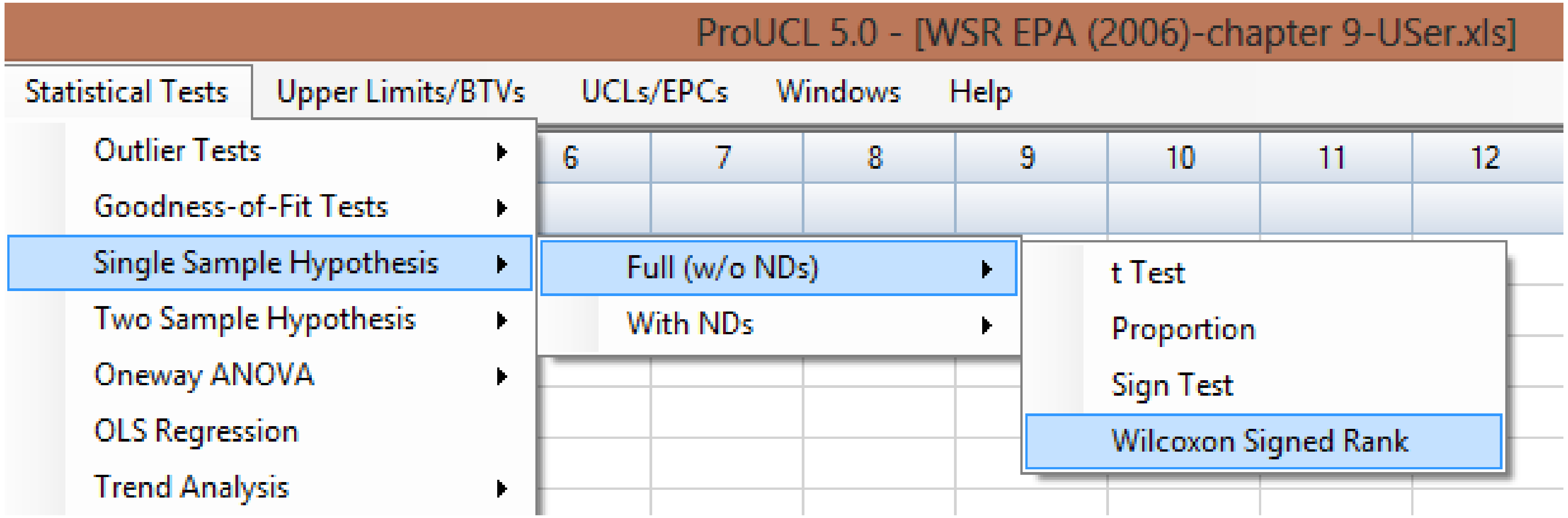


1. State the null hypothesis H_0
2. State the alternative hypothesis H_A
3. Set confidence level $1-\alpha$
4. Collect data
5. Calculate a test statistic
6. Construct acceptance/rejection region
7. Based on steps 5 and 6, draw a conclusion about H_0

Single sample hypothesis testing

- *One sample t-test*
  - Assumes normality of data set
  - Can’t be used for censored data
  - Large data set required depending on the data skewness
- *One-Sample Sign Test or Wilcoxon Signed Rank (WSR) Test*
  - Can handle NDs
  - Requires ND < C_s

- *Percentile Test*
  - to compare exceedances to the actionable level
  - Can handle NDs
  - Requires ND < C_s



Single sample hypothesis testing

- Ground water data
  - Is Cu concentration lower than XX?
  - Is Zn concentration higher than YY?

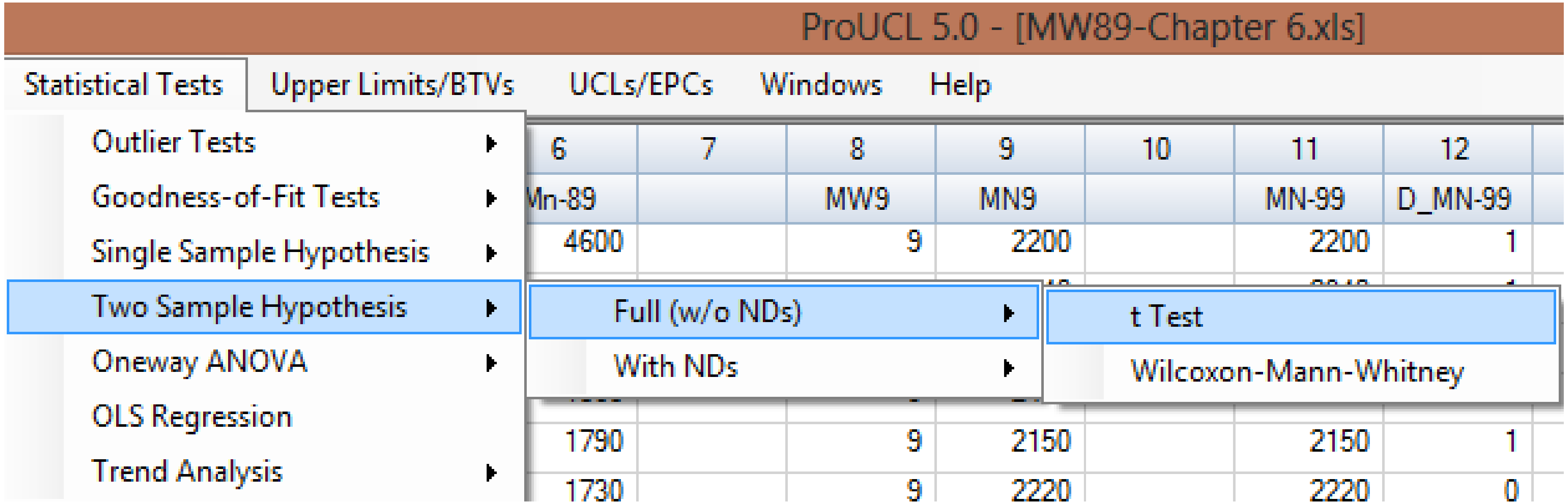
Two-sample hypothesis testing

Without NDs

- Student’s t and Satterthwaite tests
  - to compare the means of two populations (e.g. Background versus AOC).
- F-test
  - to the check the equality of dispersions of two populations.
- Two-sample nonparametric Wilcoxon-Mann-Whitney (WMW) test
  - equivalent to Wilcoxon Rank Sum (WRS) test

With NDs

- Wilcoxon-Mann-Whitney test
  - All observations (including detected values) below the highest detection limit are treated as ND (less than the highest DL) values
- Gehan’s test and Tarone-Ware test
  - useful when multiple detection limits may be present



Two sample hypothesis testing

- Groundwater data
  - Is concentration of Cu equal in Alluvial Fan and Basin Trough?
  - Is Zn concentration greater in Alluvial Fan than in Basin Trough?

Final remarks

- Take time to carefully prepare and organize data
- When in doubt consult statistician
- Don’t be quick to discard the data
    - You need to have a good scientifically justified reason
- Document well steps of analysis and decisions you make

Next ProUCL Webminars

ProUCL Utilization 2020: Part 2: Trend Analysis

Feb 10, 2020

1:00PM-2:30PM EST

ProUCL Utilization 2020: Part 3: Background Level Calculations

Mar 9, 2020

1:00PM-2:30PM EST

Contact Information for ProUCL

Felicia Barnett, EPA SCMTSC
barnett.felicia@epa.gov

Travis Linscome-Hatfield, Neptune and Company, Inc
travis@neptuneinc.org

Polona Carson, Neptune and Company, Inc
pcarson@neptuneinc.org