NUREG-1569 Statistical Methods Workshop

Dr. Alex Sun
CNWRA
Southwest Research Institute®

August 26,2010

CNWRA
A center of excellence in earth sciences and engineering®

Outline

- Statistics Refresher
- Establishment of Baseline Water Quality
- Outlier Assessment
- Restoration Target Assessment
- Trend Analysis
- Contaminant Hot Spot Evaluation

CNWRA
A center of excellence in earth sciences and engineering®

Statistics Refresher

CNWRA
A center of excellence in earth sciences and engineering®

Basic Concepts

- Sample Space: the set of all possible outcomes of a random experiment
- Sample: a finite subset of the sample space
- Random Variable (RV): a function that assigns a real number to each outcome in the sample space
- Statistic: a function of observations in a random sample, which is a RV itself

$$Y = f(x) + \varepsilon$$

CNWRA
A center of excellence in earth sciences and engineering®

Basic Concepts (cont'd)

- Sample Distribution: the probability distribution of a statistic
- Point Estimate: the numerical estimate of a probability distribution parameter
- Unbiased Point Estimator $E(\hat{\Theta}) = \theta$
- Mean Square Error of a Point Estimate

$$MSE = E\left(\hat{\Theta} - \theta\right)^2 = E\left(\hat{\Theta} - E(\hat{\Theta})\right)^2 + (\theta - E(\hat{\Theta}))^2$$

- Statistical Inference: the process of reasoning from a sample to conclusions for a population
  - Parameter estimation
  - Hypothesis testing

CNWRA
A center of excellence in earth sciences and engineering®

Point Estimators

- Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Population variance

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

- Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

CNWRA
A center of excellence in earth sciences and engineering®

Point Estimators (cont'd)

- Sample Autocorrelation

$$r = \frac{\sum_{i=1}^{n-|h|}(x_i - \bar{x})(x_{i+|h|} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Lag=1

$$r = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

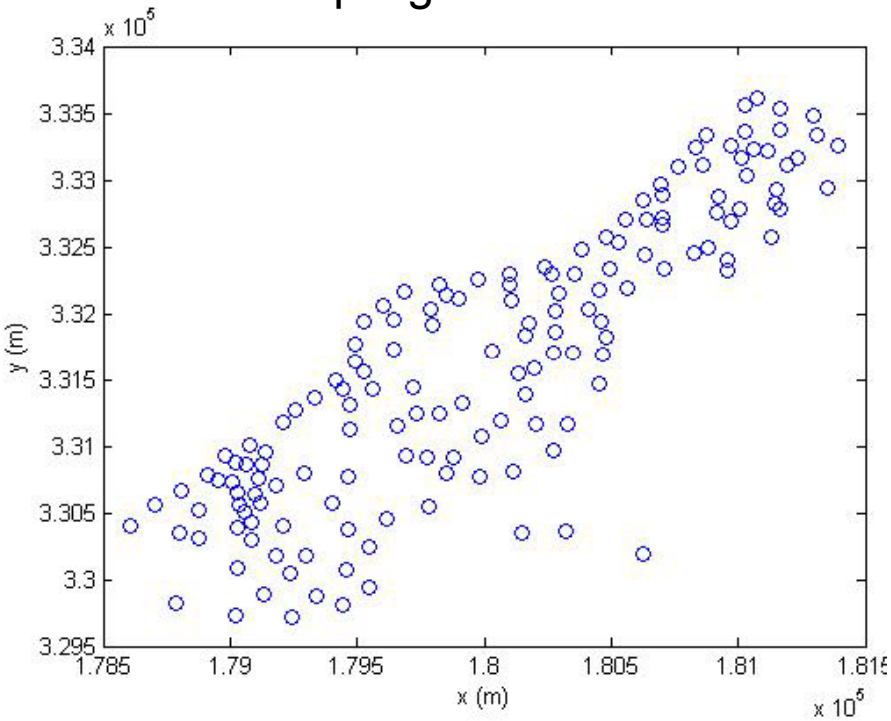
- Spatial Autocorrelation (semivariogram)

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}(x_i - y_i)^2$$

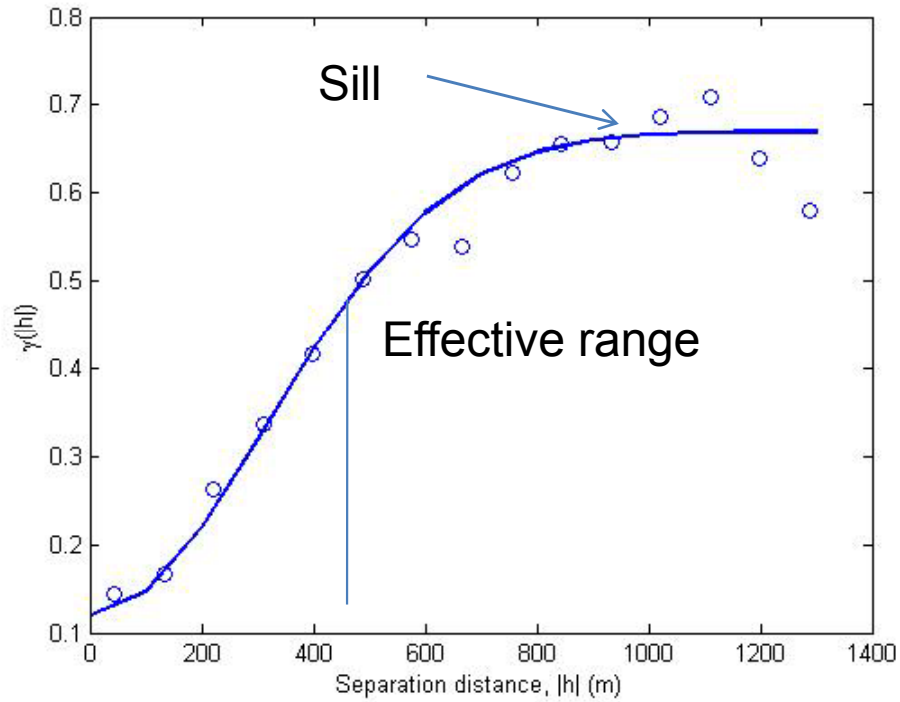
$\mathbf{x}_i$, $\mathbf{y}_i$ separated by h

Example: Spatial analysis

Sampling Locations of Zn



Semivariogram



Fitted Gaussian variogram with nugget 0.12, effective range 450, sill value (variance) 0.55

$$\gamma(h) = c\left\{1 - \exp\left[-\left(\frac{3h}{a}\right)^2\right]\right\}$$

CNWRA
A center of excellence in earth sciences and engineering®

Common Assumptions

- Statistical independence of samples
- Temporal and/or spatial stationarity
- Lack of statistical outliers
- Certain probabilistic distribution

CNWRA
A center of excellence in earth sciences and engineering®

Stationarity

- Stationarity: A statistical distribution that does not change over time and/or space

- Weak stationarity (2nd-order stationarity):
  - Mean and variance are constant
  - Autocovariance is a function of Δt for temporal or separation distance for spatial

CNWRA
A center of excellence in earth sciences and engineering®

Statistical Independence

$$P(A \mid B) = P(A), \qquad P(A \cap B) = P(A)P(B)$$

- Truly random sampling of a single, fixed, stationary population will yield independent identically distributed (i.i.d.) samples
- It's hard to verify i.i.d. directly
- However, non-zero pairwise correlation indicates data dependence
  - Temporal correlation
  - Spatial correlation

CNWRA
A center of excellence in earth sciences and engineering®

Confidence Interval

- A statistical interval for a point estimate

Significance level or Type I error

$$P(L \leq \theta \leq U) = 1 - \alpha$$

- Confidence interval [L,U] brackets the true parameter with confidence 100(1-α)

CNWRA
A center of excellence in earth sciences and engineering®

Sample Size

- For normal distribution, the length of a 2-sided confidence interval is

Standard normal $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$$2 z_{\alpha/2} \sigma / \sqrt{n}$$

- For error $E = |\bar{x} - \mu|$ to be less than or equal to with confidence 100(1-α), sample size should be

$z_{\alpha/2} \sigma / \sqrt{n}$

Sample size = $\left(\dfrac{z_{\alpha/2} \sigma}{E} \right)^2$

Assume 100(1-α)=95% => $z_{\alpha/2} = 1.96$

If E = σ => n = 4

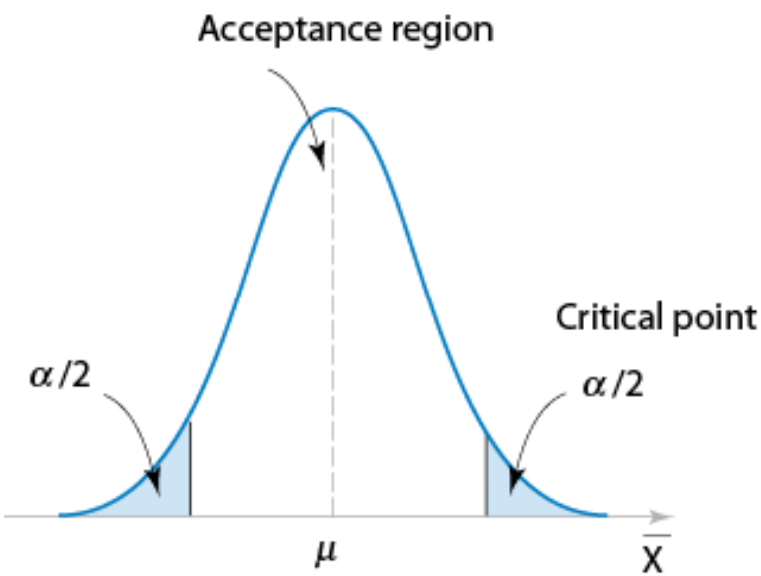
CNWRA
A center of excellence in earth sciences and engineering®

Hypothesis Testing

- Hypothesis: a statement about parameters of one or more populations
- A formal comparison is made between two mutually exclusive possible statements about reality
- Null hypothesis (H_0)
- Alternative hypothesis (H_A)
- Rejecting H_0 is a strong conclusion
- Type I error (α): Reject H_0 when H_0 is true
- Type II error (β): Fail to reject H_0 when H_A is true

Hypothesis Testing (cont'd)

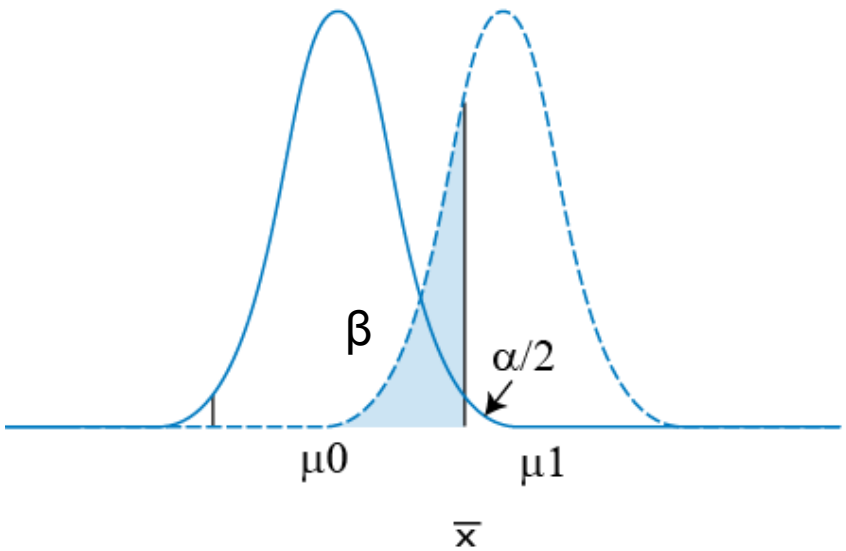
- Main steps include
  - Specify a significance level α
  - Determine an appropriate test statistic and calculate its value using a random sample
  - Decide whether or not to reject H_0.



A two-tailed test

Hypothesis Testing (cont'd)

- Type I and Type II errors are related
  - A decrease in one type of error will lead to the increase in the other type of error for fixed sample size
  - An increase in sample size will reduce both α and β
- Statistical power (1-β): the probability to reject H_0 when it is incorrect



CNWRA
A center of excellence in earth sciences and engineering®

Sample Size Revisited

- For two-sided hypothesis testing

$$n \cong \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{E^2}$$

$$E = \hat{\theta} - \theta$$

- If α=0.05, β=0.1, and *E*=σ

$$n \cong (z_{\alpha/2} + z_{\beta})^2 = (1.96 + 1.28)^2 \approx 11$$

- Conversely if *n*=4, α=0.05, *E*=σ, then β=0.516

So statistical power decreases from 0.9 to 0.48

CNWRA
A center of excellence in earth sciences and engineering®

Hypothesis Testing vs. Confidence Interval

- Confidence interval provides a range of likely values for point estimate
- Hypothesis testing provides a framework for displaying risk levels associated with a decision
- For example, if [L,U] is 100(1- α) confidence interval and H_0: $\theta=\theta_0$

  =>H_0 will be rejected if and only if θ_0 is not in [L,U]

CNWRA
A center of excellence in earth sciences and engineering®

Exploratory Tools

- Time series plot
- Box plot
- Histogram
- Autocovariance
- Variogram

CNWRA
A center of excellence in earth sciences and engineering®

Goodness of Fit Test of Normality

- Shapiro-Wilk (Shapiro-Francia) Test
  - Considered one of the most powerful methods
  - Ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance
- Kolmogorov-Smirnov Test
  - The distance between the distribution function F(x) and the empirical distribution function Fn(x)
  - Not as powerful as the Shapiro-Wilk test
  - For large sample sizes (> 50)

CNWRA
A center of excellence in earth sciences and engineering®

1.Establishement of Baseline Water Quality

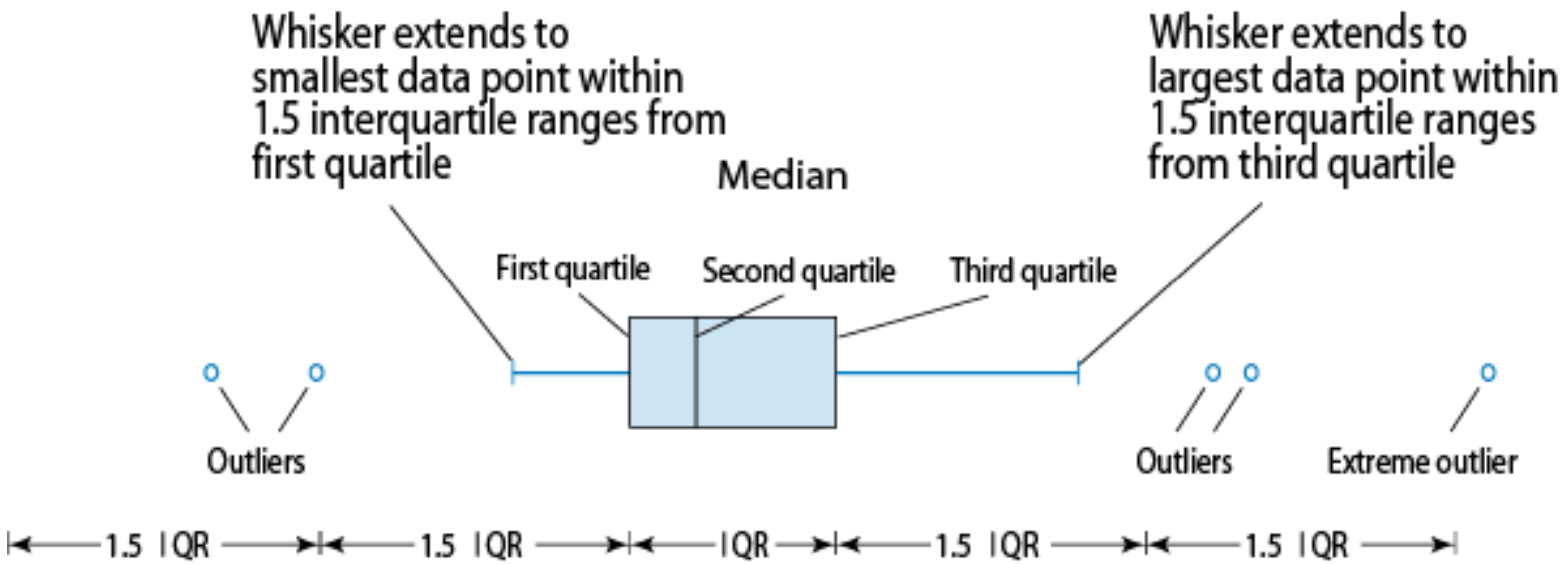
CNWRA
A center of excellence in earth sciences and engineering®

Review Procedure

- *Evaluate the sufficiency of sample size and sampling frequency (EPA suggests n> 8 to 10)*
- Examine sample statistics by applying common exploratory tools such as time series plots, box plots, histograms, scatter plots, and probability plots
- *Determine whether observations designated as baseline still reflect the current site conditions such as groundwater flow characteristics*
- Determine whether spatial and temporal trends exist in the baseline water data
- Outlier and trend analysis
- Determine baseline statistics

CNWRA
A center of excellence in earth sciences and engineering®

Box Plot



CNWRA
A center of excellence in earth sciences and engineering®

One-Way ANOVA

- Test for spatial variability
- Null hypothesis: the mean concentrations of multiple groups are equal
- Main assumptions underlying ANOVA
  - Observations are i.i.d.
  - Residuals of each group are normal
  - Variances of groups are similar

CNWRA
A center of excellence in earth sciences and engineering®

Before Applying ANOVA

- Test normality
  - Probability plot
  - Shapiro-Wilk's test
- Test the assumption of equal variance
  - Side-by-side box plot
  - Levene's test

CNWRA
A center of excellence in earth sciences and engineering®

Levene's Test

- Formal test for equal variance. Test statistic

$$z_{ij} = \left| x_{ij} - \overline{x}_i \right|$$

- Run a one-way ANOVA on z_{ij}

- Null hypothesis: all groups have equal variance

CNWRA
A center of excellence in earth sciences and engineering®

Kruskal-Wallis Test

- Non-parametric
- Test equality of medians among groups
- Does not assume normal distribution
- Identical to ANOVA but replace data with their ranks
- Assume equal variance among groups
  - Use Levene's test

CNWRA
A center of excellence in earth sciences and engineering®

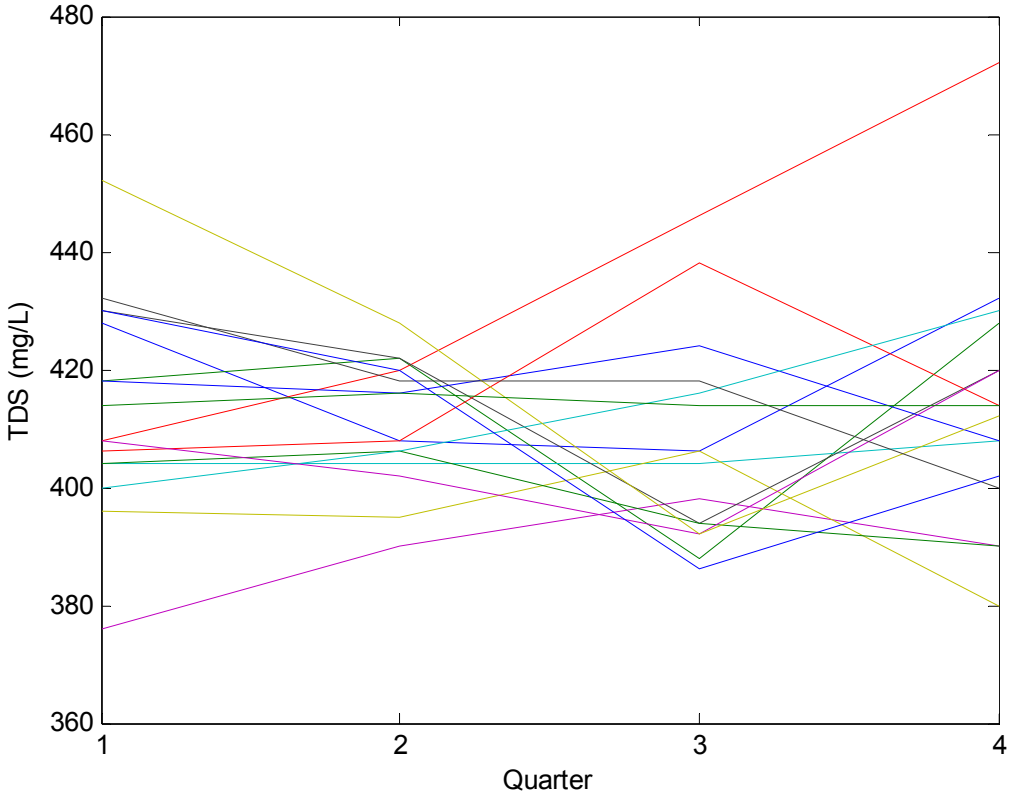
Example 1.1

Baseline TDS (mg/L)

	Quarter			
Well	*1*	*2*	*3*	*4*
3D12-1	428	408	406	432
3G13-1	418	422	388	428
3J26-1	408	420	446	472
3L16-1	404	404	404	408
3L20-1	376	390	398	390
3N25-1	396	395	406	380
3O37-2	430	422	394	420
3Q22-1	418	416	424	408
3S20-1	404	406	394	390
3T27-2	406	408	438	414
3T37-1	400	406	416	430
3U45-1	408	402	392	420
3V52-2	452	428	392	412
3V58-2	432	418	418	400
3W67-1	430	420	386	402
3AA57-1	414	416	414	414

CNWRA
A center of excellence in earth sciences and engineering®

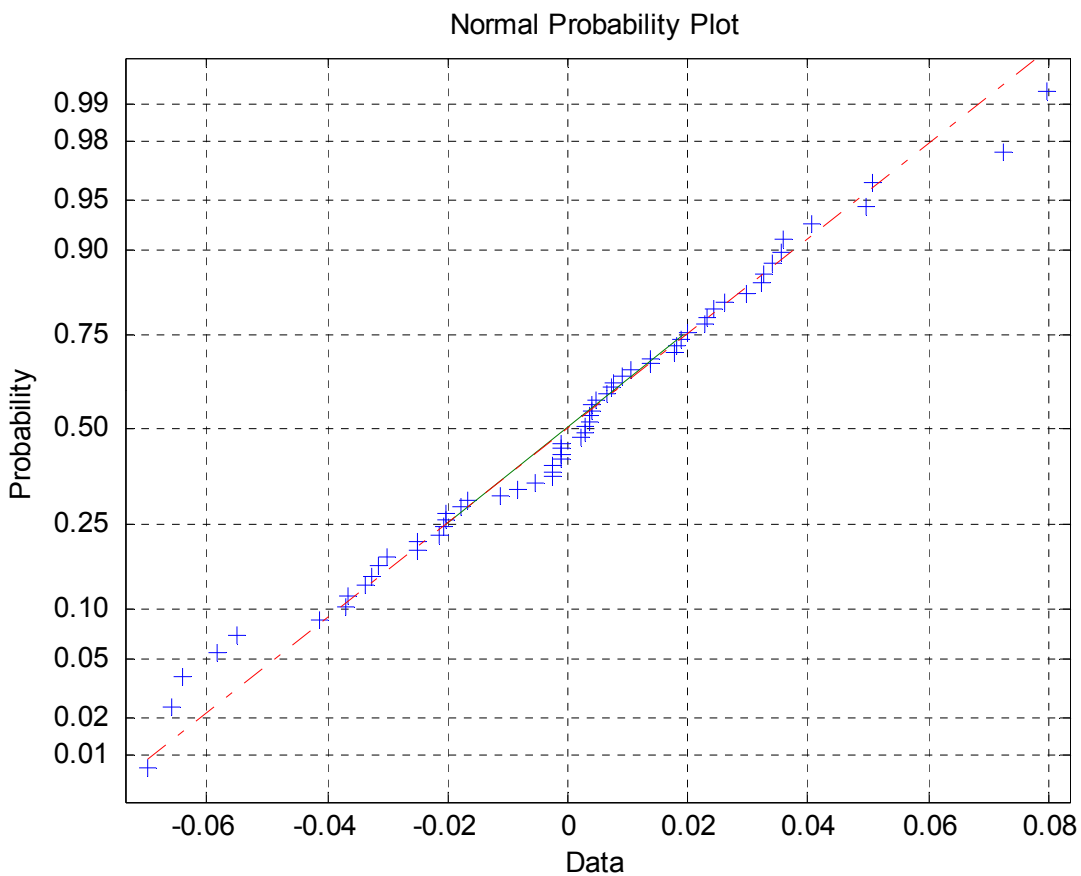
Example 1.1:Time Series Plot



CNWRA
A center of excellence in earth sciences and engineering®

Example 1.1: Probability Plot

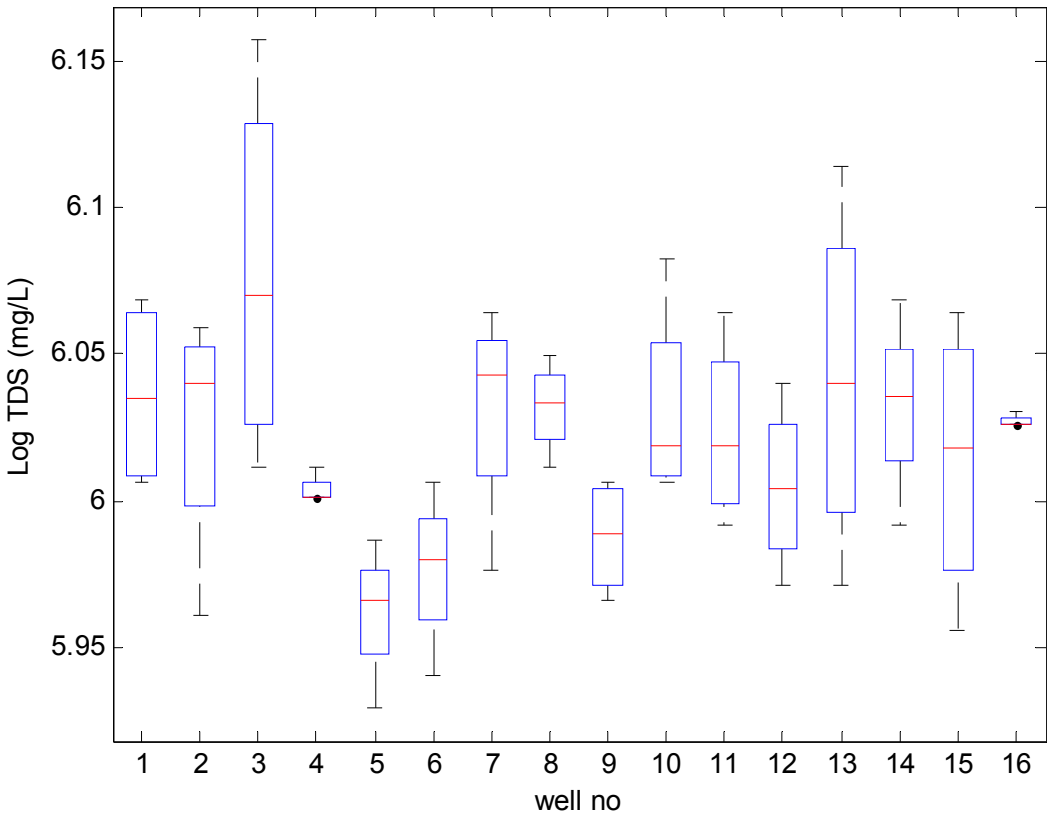
Test of normality on pooled residuals



Shapiro-Wilk test at 0.05 significance level indicates normal residuals

CNWRA
A center of excellence in earth sciences and engineering®

Example 1.1: Box Plot



CNWRA
A center of excellence in earth sciences and engineering®

Example 1.1: Levene's Test

	Quarter			
Well	*1*	*2*	*3*	*4*
3D12-1	0.0228	0.025	0.0299	0.0321
3G13-1	0.0103	0.0199	0.0641	0.034
3J26-1	0.0659	0.037	0.0231	0.0798
3L16-1	0.0025	0.0025	0.0025	0.0074
3L20-1	0.0325	0.0041	0.0244	0.0041
3N25-1	0.0047	0.0022	0.0296	0.0365
3O37-2	0.0324	0.0137	0.055	0.0089
3Q22-1	0.0037	0.0011	0.0179	0.0205
3S20-1	0.0138	0.0188	0.0112	0.0214
3T27-2	0.0251	0.0202	0.0508	0.0056
3T37-1	0.0316	0.0167	0.0076	0.0407
3U45-1	0.0065	0.0084	0.0335	0.0354
3V52-2	0.0724	0.0179	0.07	0.0202
3V58-2	0.0357	0.0028	0.0028	0.0412
3W67-1	0.0497	0.0262	0.0582	0.0176
3AA57-1	0.0012	0.0036	0.0012	0.0012

Levene's test: not reject null hypothesis of equal variance at 0.01 level

CNWRA
A center of excellence in earth sciences and engineering®

Example 1.1: One-Way ANOVA

- *p* wells, *n* total samples
- Treat each well as a group

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	*F*-Statistic
Between Wells	SS_{wells}	$p-1$	$MS_{wells} = SS_{wells}/(p-1)$	$F = MS_{wells}/MS_{error}$
Error (within wells)	SS_{error}	$n-p$	$MS_{error} = SS_{error}/(n-p)$	
Total	SS_{total}	$n-1$		

$$SS_{wells} = \sum_{i=1}^{p} n_i \left(\bar{x}_i - \hat{x}\right)^2$$

$$SS_{err} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

CNWRA
A center of excellence in earth sciences and engineering®

Example 1.1: One-Way ANOVA

One-way ANOVA Summary Table

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	0.04573	15	0.00305	2.38	0.0119
Error	0.06156	48	0.00128		
Total	0.10728	63			

Not significant at α=0.01
Natural variability is statistically insignificant

Example 1.1 Kruskal-Wallis

- Calculate Kruskal-Wallis test statistic

Group mean

Global mean rank=(n+1)/2

$$K = (n-1)\frac{\sum_{i=1}^{p} n_i(\bar{r}_i - \hat{r})^2}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(r_{ij} - \hat{r})^2}$$

- p = 0.035, not significant at α =0.01 level

CNWRA
A center of excellence in earth sciences and engineering®

Issue

- Interwell mean (e.g., wellfield averaging) is not appropriate when significant spatial variability exists
- Intrawell prediction limits then become more appropriate
- However, with only 4 samples per baseline well the statistical power of intrawell test is low

CNWRA
A center of excellence in earth sciences and engineering®

Prediction Limit Primer

- A prediction interval for each constituent is established from the distribution of the background data, and the level of each constituent in each compliance well is compared to the prediction interval
- Parametric prediction limit

$$PL = \bar{x} + t_{1-\alpha,n-1} s \sqrt{1 + 1/n}$$

100(1-α) is the confidence interval associated with PL

CNWRA
A center of excellence in earth sciences and engineering®

Improve Power of Prediction Interval

- Pooling variance estimates: Substitute intrawell standard deviation using RMSE (i.e., square root of MS_{error})
- RMSE provides a better estimate of the typical intrawell variation because it is based on levels of fluctuation averaged across several wells

$$PL = \overline{x} + t_{1-\alpha, p*(n-1)} \sqrt{MS_{err}(1 + 1/n)}$$

The degree of freedom (d.o.f) of MS_{err} is p*(n-1)

CNWRA
A center of excellence in earth sciences and engineering®

Example 1.2

- Use baseline TDS data in Example 1.1
- n = 4, p=16
- MS_{err} = 0.00128
- D.o.f. for improved method: (4-1)*16=48
- At α=0.01,
  - $t_{1-\alpha,4}$=4.5407
  - $t_{1-\alpha,48}$=2.4066

CNWRA
A center of excellence in earth sciences and engineering®

Example 1.2 (cont'd)

	Quarter					
Well	*1*	*2*	*3*	*4*	***PL***	***PL-Boost***
3D12-1	428	408	406	432	492.1886	460.6118
3G13-1	418	422	388	428	516.8687	455.5097
3J26-1	408	420	446	472	605.6833	479.8474
3L16-1	404	404	404	408	415.2523	445.9207
3L20-1	376	390	398	390	438.0429	427.6679
3N25-1	396	395	406	380	452.7985	433.9674
3O37-2	430	422	394	420	504.9707	458.3405
3Q22-1	418	416	424	408	451.4605	458.5435
3S20-1	404	406	394	390	439.6873	438.7061
3T27-2	406	408	438	414	496.9073	458.3759
3T37-1	400	406	416	430	484.6619	454.5624
3U45-1	408	402	392	420	469.2636	446.3361
3V52-2	452	428	392	412	570.7233	462.9115
3V58-2	432	418	418	400	489.332	458.9662
3W67-1	430	420	386	402	521.5578	450.4944
3AA57-1	414	416	414	414	419.6008	456.3838

CNWRA
A center of excellence in earth sciences and engineering®

2. Restoration Target Assessment

CNWRA
A center of excellence in earth sciences and engineering®

Review Methods

- Hypothesis testing
  - EPA recommends confidence interval testing over fixed limit (e.g., restoration standard) testing
  - Confidence intervals explicitly account for variation and uncertainty in the sample data used to construct them
  - Use upper confidence interval for restoration for target assessments
  - H_0: constituent concentration >= restoration standard
  - H_1: constituent concentration < restoration standard
- One-sample testing
- Two-sample testing
- Multiple sample testing
  - One-way ANOVA

CNWRA
A center of excellence in earth sciences and engineering®

The *P*-Value

- *P*-value is the smallest level of significance that would lead to rejection of the null hypothesis
- It allows the decisionmaker to decide significance level
- For normal distribution, two-tailed test

$$P = 2\left[1 - \Phi(|z_0|)\right]$$

CNWRA
A center of excellence in earth sciences and engineering®

P-Value (cont'd)

From Example 1.1

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	0.04573	15	0.00305	2.38	0.0119
Error	0.06156	48	0.00128		
Total	0.10728	63			

Significant at α = 0.05,
Not significant at α = 0.01

CNWRA
A center of excellence in earth sciences and engineering®

One-Sample Tests

- Test on the mean of normal distribution w/ known variance: standard normal
- Test on the mean of normal distribution w/ unknown variance: Student's *t*
- Test on the variance of normal distribution: Chi-square

CNWRA
A center of excellence in earth sciences and engineering®

Two-Sample Tests

- Compare statistics between two samples
  - For example, compare restoration monitoring (stability monitoring) to baseline samples
- T-test: Test for equality of two means
  - assumes equal variance and normal distribution

$$t = \left(\bar{x}_c - \bar{x}_{bg}\right) \Big/ \sqrt{\left[\underbrace{\frac{(n_{bg}-1)s_{bg}^2 + (n_c-1)s_c^2}{(n_{bg}+n_c-2)}}_{\text{Pooled estimator of variance}}\right]\left(\frac{1}{n_{bg}} + \frac{1}{n_c}\right)}$$

CNWRA
A center of excellence in earth sciences and engineering®

Two-Sample Tests (cont'd)

- Welch's T-test: Test for equality of two means
  - For unequal variances
  - Should be used for most of T-tests

$$t = \left(\overline{x}_c - \overline{x}_{bg}\right) \Big/ \sqrt{\frac{s_{bg}^2}{n_{bg}} + \frac{s_c^2}{n_c}}$$

- Wilcoxon Rank-Sum test: Test for medians
- Ratio between two chi-square random variables: *F* distribution

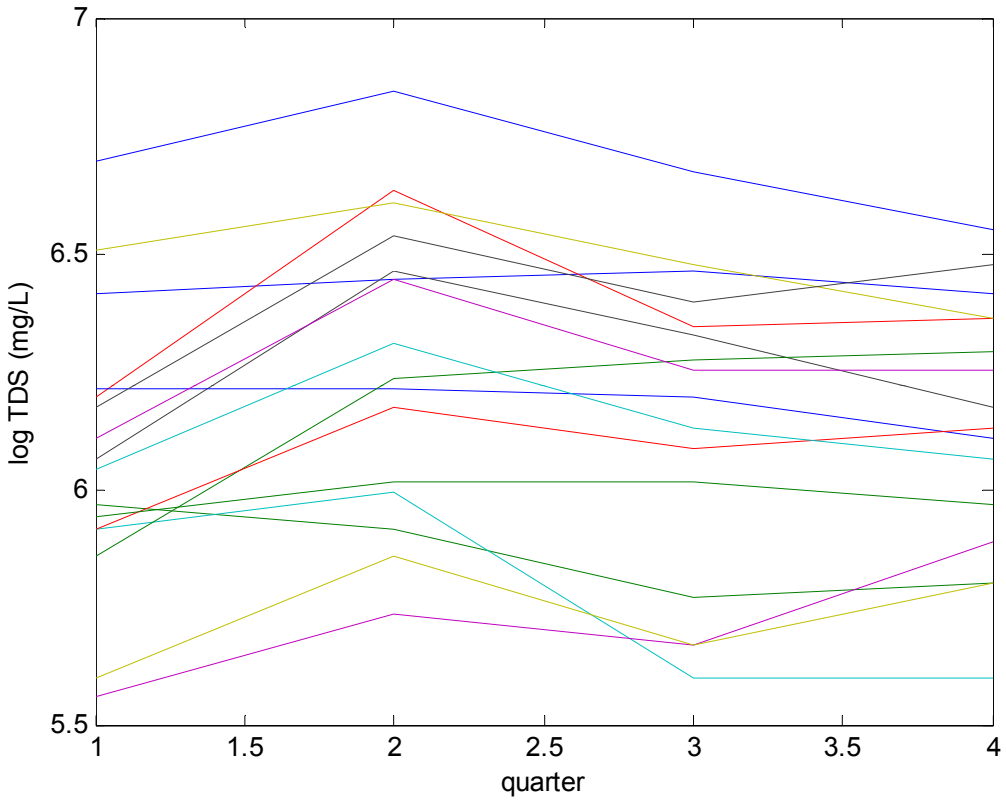
CNWRA
A center of excellence in earth sciences and engineering®

Example 2.1

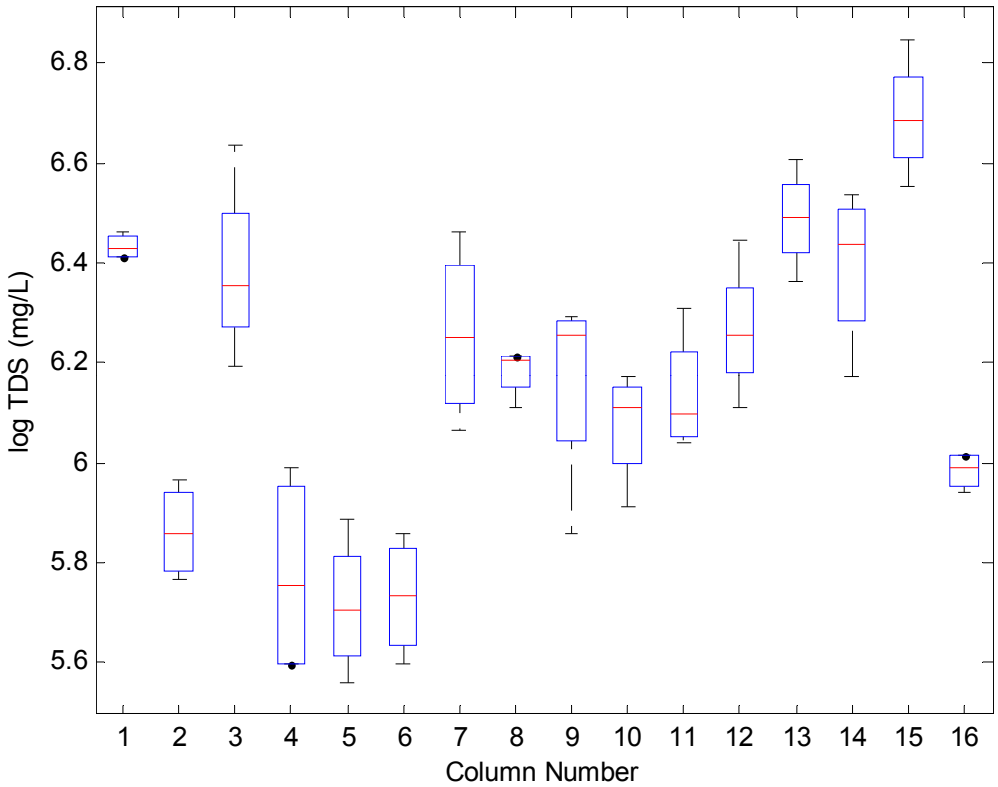
TDS stability monitoring data (mg/L)

	Quarter			
Well	*1*	*2*	*3*	*4*
3D12-1	610	630	640	610
3G13-1	390	370	320	330
3J26-1	490	760	570	580
3L16-1	370	400	270	270
3L20-1	260	310	290	360
3N25-1	270	350	290	330
3O37-2	430	640	560	480
3Q22-1	500	500	490	450
3S20-1	350	510	530	540
3T27-2	370	480	440	460
3T37-1	420	550	460	430
3U45-1	450	630	520	520
3V52-2	670	740	650	580
3V58-2	480	690	600	650
3W67-1	810	940	790	700
3AA57-1	380	410	410	390

Example 2.1 (cont'd)



Example 2.1 (cont'd)



Levene's test: Not to reject H_0
Shapiro-Wilk's test: Not to reject H_0 ⇨ Apply ANOVA

CNWRA
A center of excellence in earth sciences and engineering®

Example 2.1 (cont'd)

UCL approach: Upper 95% confidence limit should be below the TDS MCL 500 mg/L

	Quarter				
Well	*1*	*2*	*3*	*4*	*UCL*
3D12-1	610	630	640	610	640.2269
3G13-1	390	370	320	330	392.2097
3J26-1	490	760	570	580	734.3218
3L16-1	370	400	270	270	411.1202
3L20-1	260	310	290	360	355.4809
3N25-1	270	350	290	330	354.4827
3O37-2	430	640	560	480	640.2047
3Q22-1	500	500	490	450	514.0473
3S20-1	350	510	530	540	605.5118
3T27-2	370	480	440	460	498.0565
3T37-1	420	550	460	430	533.7094
3U45-1	450	630	520	520	618.9888
3V52-2	670	740	650	580	739.9027
3V58-2	480	690	600	650	722.7392
3W67-1	810	940	790	700	928.8488
3AA57-1	380	410	410	390	415.3985

CNWRA
A center of excellence in earth sciences and engineering®

Example 2.1 (cont'd)

Welch's T-test

Stability monitoring						Baseline				
Well	*1*	*2*	*3*	*4*	*H0*	*Well*	*1*	*2*	*3*	*4*
3D12-1	610	630	640	610	0	3D12-1	428	408	406	432
3G13-1	390	370	320	330	1	3G13-1	418	422	388	428
3J26-1	490	760	570	580	0	3J26-1	408	420	446	472
3L16-1	370	400	270	270	0	3L16-1	404	404	404	408
3L20-1	260	310	290	360	1	3L20-1	376	390	398	390
3N25-1	270	350	290	330	1	3N25-1	396	395	406	380
3O37-2	430	640	560	480	0	3O37-2	430	422	394	420
3Q22-1	500	500	490	450	0	3Q22-1	418	416	424	408
3S20-1	350	510	530	540	0	3S20-1	404	406	394	390
3T27-2	370	480	440	460	0	3T27-2	406	408	438	414
3T37-1	420	550	460	430	0	3T37-1	400	406	416	430
3U45-1	450	630	520	520	0	3U45-1	408	402	392	420
3V52-2	670	740	650	580	0	3V52-2	452	428	392	412
3V58-2	480	690	600	650	0	3V58-2	432	418	418	400
3W67-1	810	940	790	700	0	3W67-1	430	420	386	402
3AA57-1	380	410	410	390	0	3AA57-1	414	416	414	414

CNWRA
A center of excellence in earth sciences and engineering®

3. Outlier Analysis

CNWRA
A center of excellence in earth sciences and engineering®

Review Methods

- Box plot
- Dixon's Test
- Rosner's Test
- Skewness Test

CNWRA
A center of excellence in earth sciences and engineering®

Dixon's Test

- Test for a single outlier for sample size <=25
- Assumes that the data set minus the suspected outlier is normally distributed
- Masking: an extreme outlier being missed because one or more outliers are present

CNWRA
A center of excellence in earth sciences and engineering®

Dixon's Test

- Order the data set
- Check normality of remaining data
- For low outlier, compute test statistic

$$c = \begin{cases} (x_2 - x_1)/(x_n - x_1) & 3 \le n \le 7 \\ (x_2 - x_1)/(x_{n-1} - x_1) & 8 \le n \le 10 \\ (x_3 - x_1)/(x_{n-1} - x_1) & 11 \le n \le 13 \\ (x_3 - x_1)/(x_{n-2} - x_1) & 14 \le n \le 25 \end{cases}$$

- For high outlier, compute

$$c = \begin{cases} (x_n - x_{n-1})/(x_n - x_1) & 3 \le n \le 7 \\ (x_n - x_{n-1})/(x_n - x_2) & 8 \le n \le 10 \\ (x_n - x_{n-2})/(x_n - x_2) & 11 \le n \le 13 \\ (x_n - x_{n-2})/(x_n - x_3) & 14 \le n \le 25 \end{cases}$$

CNWRA
A center of excellence in earth sciences and engineering®

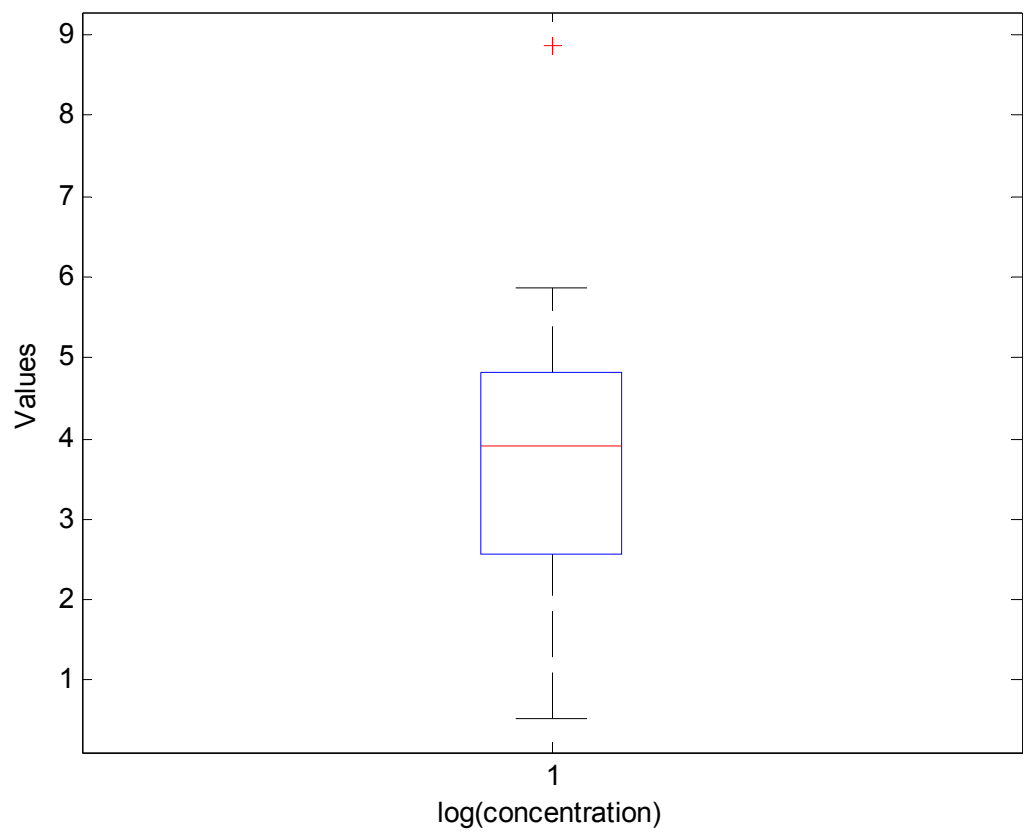
Example 3.1

Sample from 5 background wells (in ppb), n=20

	Well 1	Well 2	Well 3	Well 4	Well 5
1	1.7	302	16.2	199	275
2	3.2	35.1	7066	41.6	6.5
3	7.3	15.6	350	75.4	59.7
4	12.1	13.7	70.1	57.9	68.4

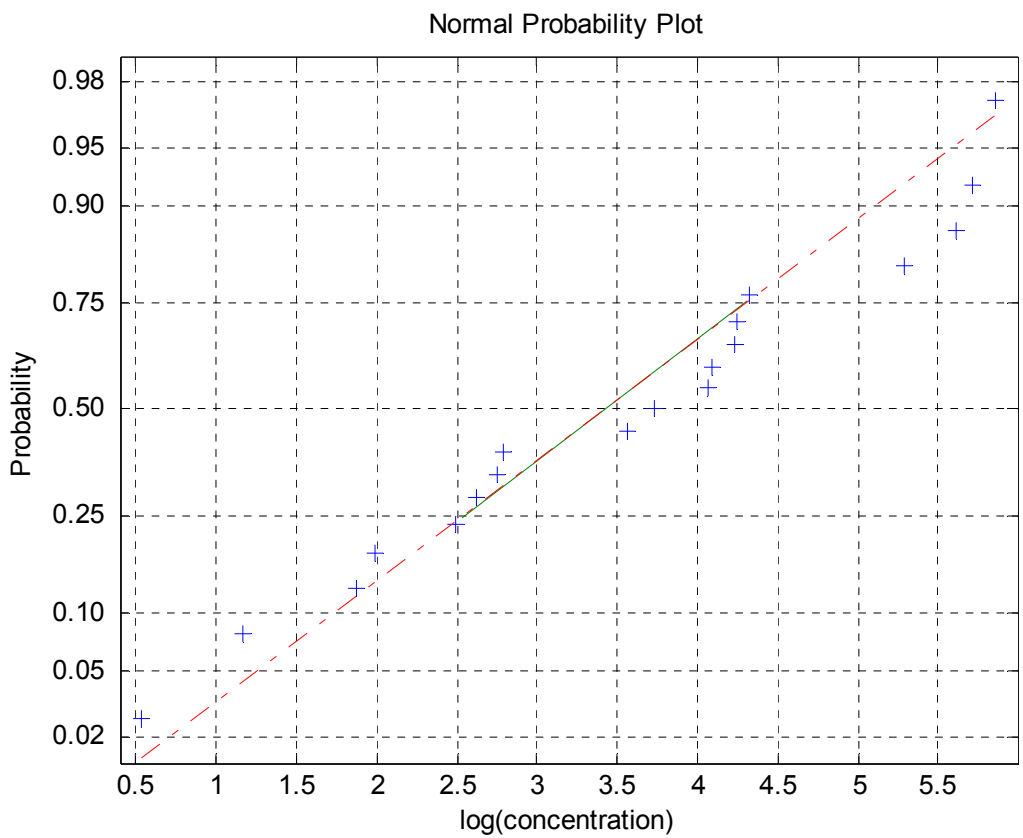
Example 3.1 (cont'd)

Box plot on log-transformed data



Example 3.1 (cont'd)

Probability plot of log-transformed data



CNWRA
A center of excellence in earth sciences and engineering®

Example 3.1 (cont'd)

- A suspected high outlier

$$
\begin{aligned}
c &= (x_n - x_{n-2})/(x_n - x_3) \\
&= (\log(7066) - \log(302))/(\log(7066) - \log(6.5)) \\
&= 0.451
\end{aligned}
$$

- With n=20, α=0.05, critical point = 0.450

- So confirm the outlier

CNWRA
A center of excellence in earth sciences and engineering®

Rosner's Test

- For sample sizes >=20
- For detecting multiple outliers
- Assumes that data without outliers follow normal distribution

CNWRA
A center of excellence in earth sciences and engineering®

Rosner's Test (cont'd)

- Order the data and identify potential number of outliers, *r*
- Form *r* subsets

$$\{\bar{x}^{(0)}, s^{(0)}, y^{(0)}\}, \{\bar{x}^{(1)}, s^{(1)}, y^{(1)}\}, \ldots\ldots, \{\bar{x}^{(r-1)}, s^{(r-1)}, y^{(r-1)}\}$$

- Calculate Rosner's test statistic

$$R_i = \frac{\left| y^{(i)} - \bar{x}^{(i)} \right|}{s^{(i)}}, \text{ for } i = r-1, \ldots, 0$$

$y^{(i)}$ i-th outlier

$\bar{x}^{(i)}$ Mean of the i-th subset

$s^{(i)}$ Standard deviation of the i-th subset

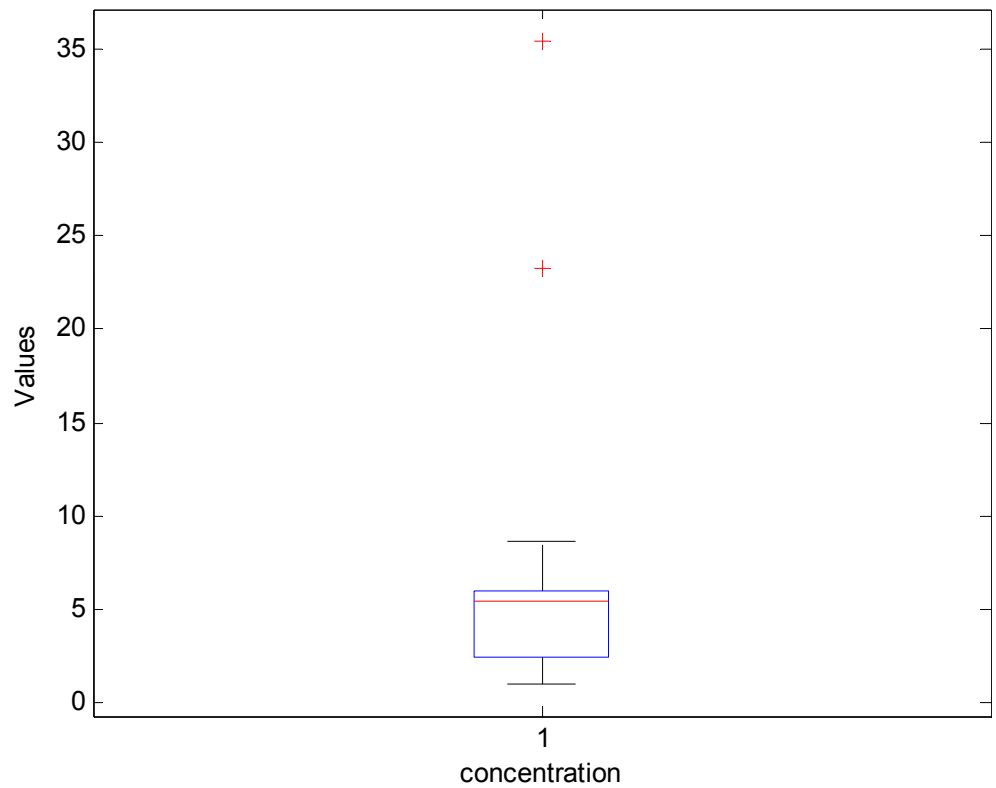
CNWRA
A center of excellence in earth sciences and engineering®

Example 3.2

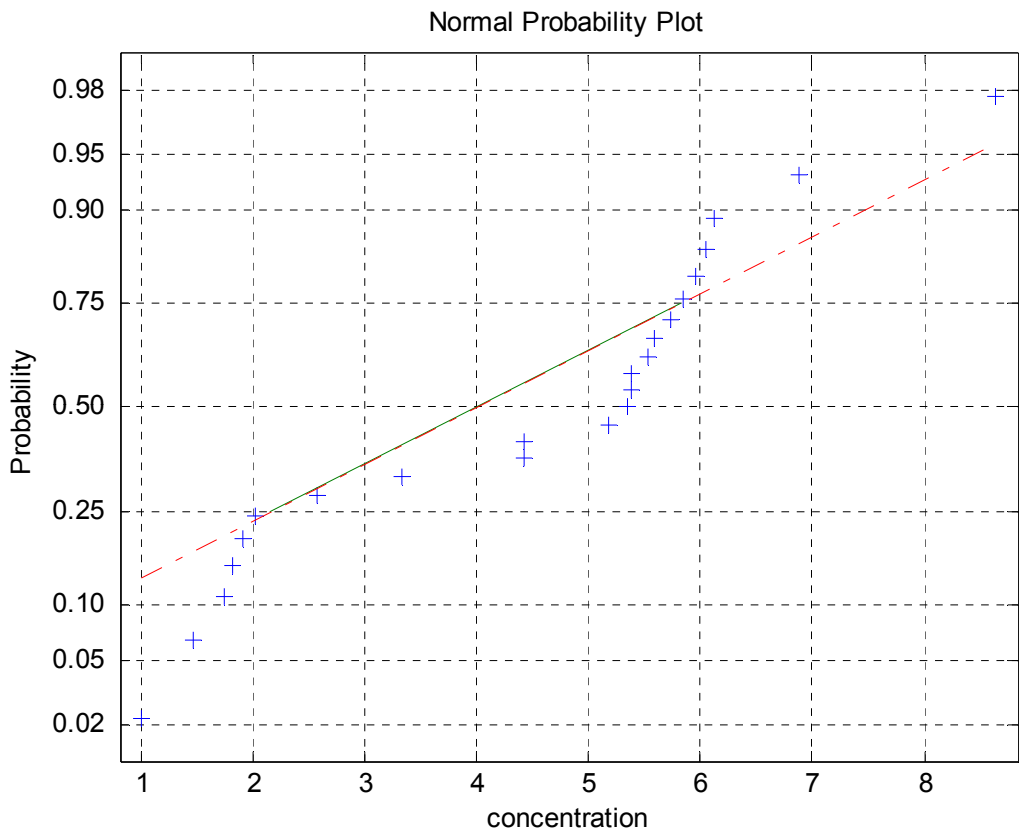
	Well 1	Well 2	Well 3	Well 4	Well 5
1	3.34	5.59	1.91	6.12	8.64
2	5.39	5.96	1.74	6.05	5.34
3	5.74	1.47	23.23	5.18	5.53
4	6.88	2.57	1.82	4.43	4.42
5	5.85	5.39	2.02	1	35.45

Example 3.2 (cont'd)

Two possible outliers



Example 3.2 (cont'd)



CNWRA
A center of excellence in earth sciences and engineering®

Example 3.2 (cont'd)

- Formulate Rosner's test data sets for r=2
- Starting with R_1
- After removing the largest suspected outlier, 35.45

Mean is 5.23, standard deviation 4.326, $y^{(0)}$=23.23,

$$R_1 = \frac{23.23 - 5.23}{4.326} = 4.16$$

- For *n*=25, 2 outliers, and α=0.05, critical point is 2.83

- So confirm the two outliers

CNWRA
A center of excellence in earth sciences and engineering®

4. Trend Analysis

CNWRA
A center of excellence in earth sciences and engineering®

Review Methods

- Linear Regression
- Mann-Kendall Test
- Sen's Test

CNWRA
A center of excellence in earth sciences and engineering®

Linear Regression

- Assumptions
  - The residuals are normal
  - Equal variance at different sampling times and for different mean concentration levels
- Slope

$$\hat{b} = \frac{\sum_{i=1}^{n}(t_i - \bar{t})x_i}{(n-1)S_t^2}$$

Mean sampling time

Variance of sampling dates

- Regression equation

$$\hat{x}_t = \bar{x} + \hat{b}(t - \bar{t})$$

CNWRA
A center of excellence in earth sciences and engineering®

Linear Regression (cont'd)

- Estimation variance

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

- Null hypothesis: slope $b = 0$

Std dev of b

$$s_{\hat{b}} = \left[s_e^2 \Big/ \sum_{i=1}^{n}(t_i - \bar{t})^2 \right]^{1/2}$$

t-statistic with (n-2) d.o.f.

$$t_{\hat{b}} = \hat{b} / s_{\hat{b}}$$

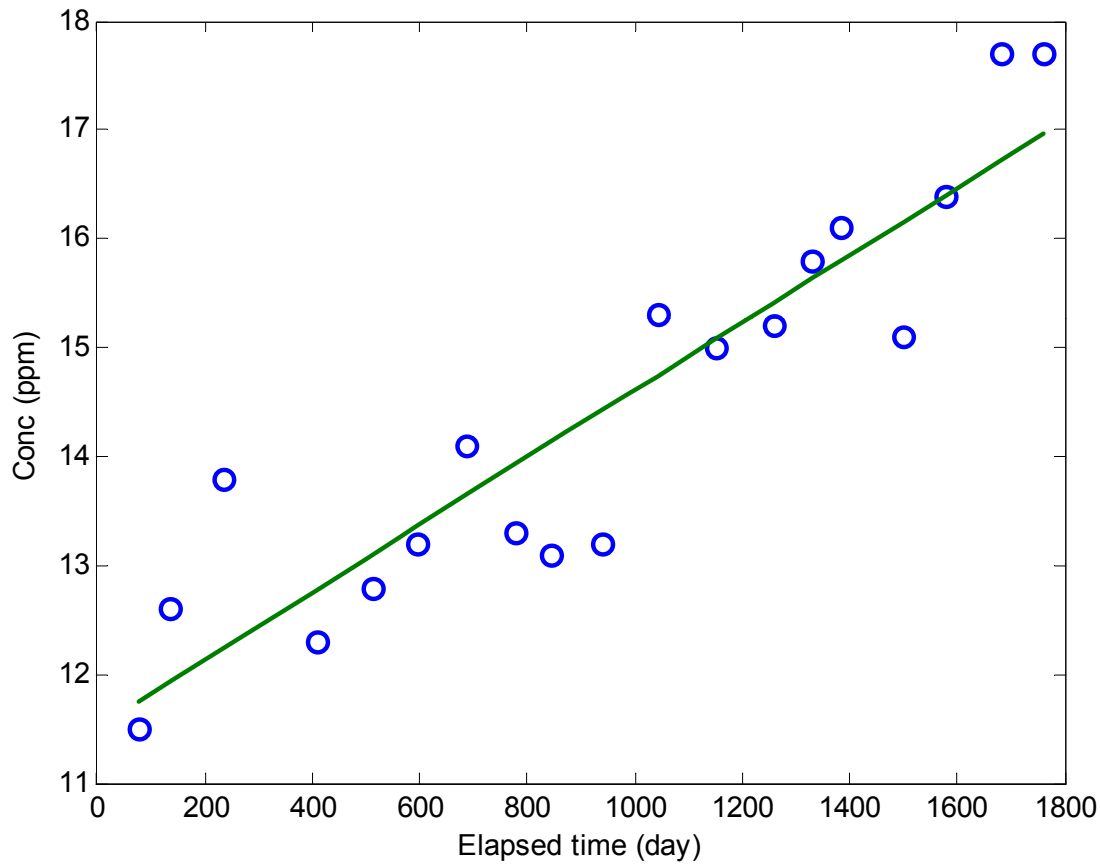
CNWRA
A center of excellence in earth sciences and engineering®

Example 4.1

Sample Date	Elapsed Days	Conc (ppm)
3/18/2002	76	11.5
5/14/2002	133	12.6
8/22/2002	233	13.8
2/12/2003	407	12.3
5/29/2003	513	12.8
8/18/2003	594	13.2
11/20/2003	688	14.1
2/19/2004	779	13.3
4/26/2004	846	13.1
7/29/2004	940	13.2
11/9/2004	1043	15.3
2/24/2005	1150	15
6/14/2005	1260	15.2
8/23/2005	1330	15.8
10/17/2005	1385	16.1
2/8/2006	1499	15.1
4/27/2006	1577	16.4
8/10/2006	1682	17.7
10/26/2006	1759	17.7

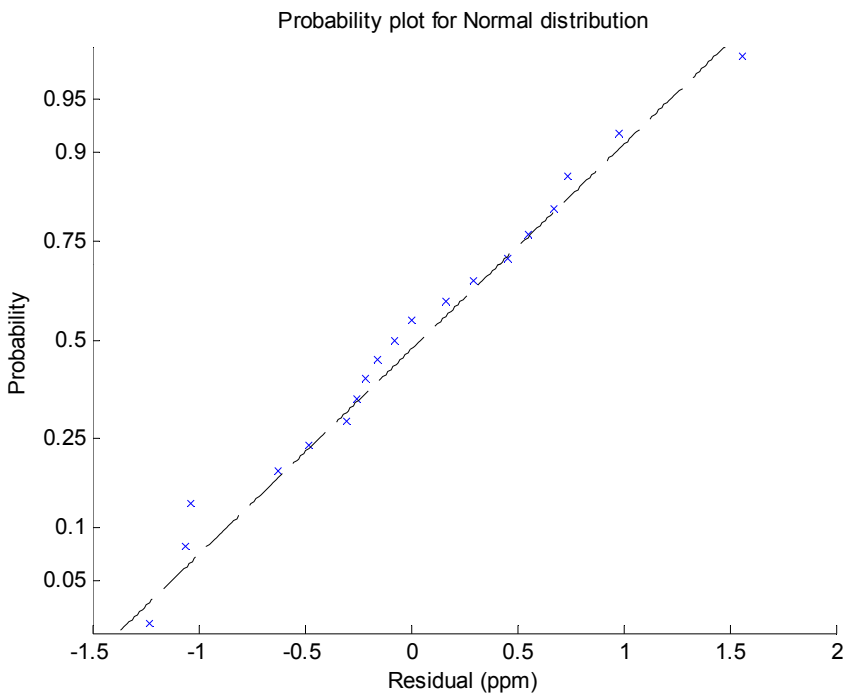
Example 4.1 (cont'd)

Linear Regression

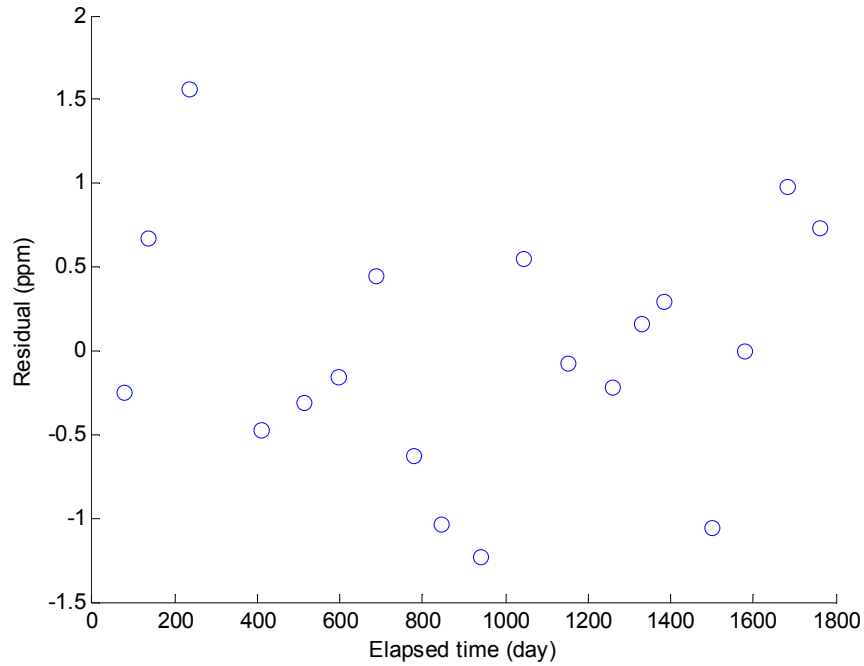


Example 4.1 (cont'd)

Probability Plot



Scatter Plot



Shapiro-Wilk Test statistic = 0.9816 > critical point = 0.9588 for α=0.01

CNWRA
A center of excellence in earth sciences and engineering®

Example 4.1 (cont'd)

- MSE = 0.5628
- Mean t = 941.79 days
- b = 0.0031
- Standard error of b

$$s_{\hat{b}} = \left[s_e^2 \Big/ \sum_{i=1}^{n} (t_i - \bar{t})^2 \right]^{1/2}$$

$$= \left[0.5628 \Big/ \left((76 - 941.79)^2 + (1759 - 941.59)^2 \right) \right]^{1/2}$$

$$= 0.00033$$

- b statistic =0.0031/0.00033=9.39 > critical value of 2.567 at α=0.01

CNWRA
A center of excellence in earth sciences and engineering®

Sen's Test

- Non-parametric test for linear trend
  - No assumption on data distribution
- Useful for estimating magnitude of increasing trend

CNWRA
A center of excellence in earth sciences and engineering®

Sen's Test (cont'd)

- Sort the data by sampling date
- Compute pairwise slope, Q

$$Q = \frac{x_i - x_j}{i - j}, i > j$$

- Total number of pairs (one measurement per event)

$$N_p = \frac{n(n-1)}{2}$$

- Find the median of all Q's

CNWRA
A center of excellence in earth sciences and engineering®

Sen's Test (cont'd)

- Variance

$$S^2 = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)\right]$$

$q =$ number of values with ties

$t_p =$ number of ties for each tied value

- Test statistic (two-sided)

$$\underbrace{\frac{N_p - Z_{1-\alpha/2}S}{2}}_{M_1} < \tilde{x} < \underbrace{\frac{N_p + Z_{1-\alpha/2}S}{2}}_{M_2}$$

normal distribution

Note: the confidence intervals are ranks

- Reject null hypothesis trend == 0 if $[Q_{M1}, Q_{M2}]$ does not contain 0

CNWRA
A center of excellence in earth sciences and engineering®

Mann-Kendall Test

- Non-parametric method for linear trend
- Nondetects and irregular sampling intervals are OK
- Valid for 40 or fewer samples

CNWRA
A center of excellence in earth sciences and engineering®

Mann-Kendall Test (cont'd)

- Sort data by sampling dates
- Calculate Mann-Kendall scores for all pairs of data ($i'>i$)

$$\mathrm{sgn}(x_{i'} - x_i) = \begin{cases} 1 & \text{if } x_{i'} - x_i > 0 \\ 0 & \text{if } x_{i'} - x_i = 0 \\ -1 & \text{if } x_{i'} - x_i < 0 \end{cases}$$

- Mann-Kendall statistic

$$S_{MK} = \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} \mathrm{sgn}(x_{i'} - x_i)$$

CNWRA
A center of excellence in earth sciences and engineering®

Mann-Kendall Test (cont'd)

- Mann-Kendall statistic variance

$$\mathrm{var}(S_{MK}) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5) \right]$$

$q =$ number of values with ties

$t_p =$ number of ties for each tied value

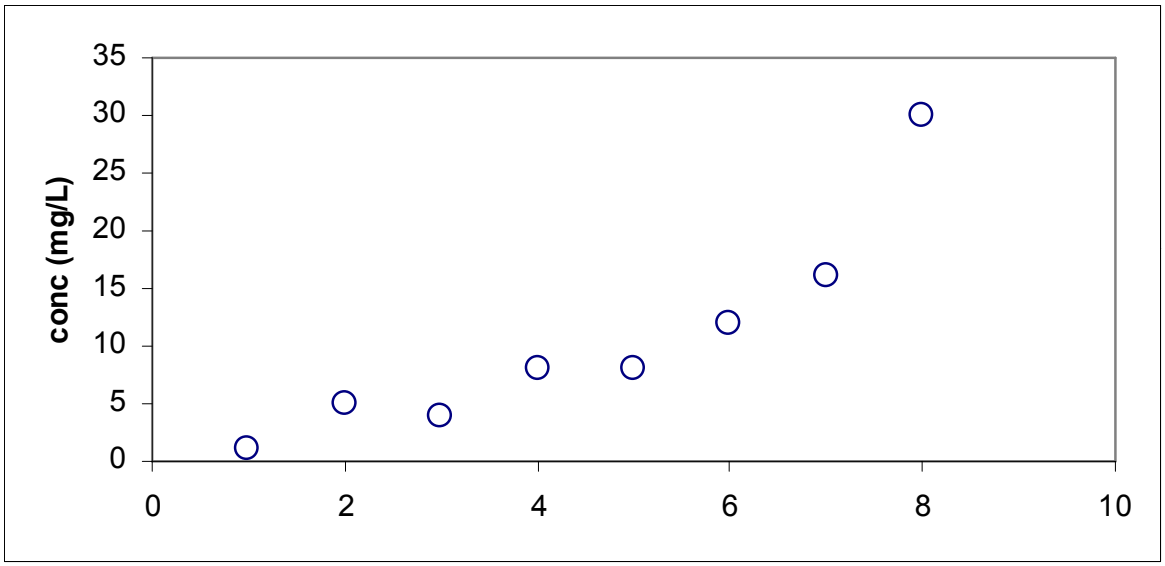
- Approximate the ratio using standard normal

$$Z = \frac{S_{MK} - 1}{\mathrm{var}(S_{MK})^{1/2}}$$

CNWRA
A center of excellence in earth sciences and engineering®

Example 4.2

Quarter	Conc (mg/L)
1	1
2	5
3	4
4	8
1	8
2	12
3	16
4	30



CNWRA
A center of excellence in earth sciences and engineering®

Example 4.2 (cont'd)

- Sen's Test
  - Calculate & rank the slopes Q
  - Median Q = 2.67
  - Var(S)=64.33
  - $[Q_{M1}, Q_{M2}]$ = [6.14, 21.86] does not contain 0; therefore, reject H_0

i	1	2	3	4	5	6	7	8
x	1	5	4	8	8	12	16	30
		4.00	1.50	2.33	1.75	2.20	2.50	4.14
			-1.00	1.50	1.00	1.75	2.20	4.17
				4.00	2.00	2.67	3.00	5.20
					0.00	2.00	2.67	5.50
						4.00	4.00	7.33
							4.00	9.00
								14.00

CNWRA
A center of excellence in earth sciences and engineering®

Example 4.2 (cont'd)

- Mann-Kendall Test
- S=26-1 = 25
- $Z=(25/64.33^{1/2})$

  $= 3.12>1.96$

- Reject null hypothesis

									Plus Sign	Minus Sign
i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**		
x	**1**	**5**	**4**	**8**	**8**	**12**	**16**	**30**		
		1	1	1	1	1	1	1	7	0
			-1	1	1	1	1	1	5	1
				1	1	1	1	1	5	0
					0	1	1	1	3	0
						1	1	1	3	0
							1	1	2	0
								1	1	0
									26	1

CNWRA
A center of excellence in earth sciences and engineering®

5. Hot spots

CNWRA
A center of excellence in earth sciences and engineering®

Review Methods

- Parallel box plots
- Parallel time series plot
- One-way ANOVA
- Skewness test
- Contaminant source analysis
- Prediction limits

CNWRA
A center of excellence in earth sciences and engineering®

Prediction Limit

- Given *n* previous measurements, prediction limit for the next <u>single</u> measurement with confidence 100(1-α)

$$PL = \bar{x} + t_{1-\alpha,n-1} s\sqrt{1+1/n}$$

- Given *n* previous measurements, prediction limit for the next <u>*k*</u> measurements

$$PL = \bar{x} + t_{1-\alpha^*/k,n-1} s\sqrt{1+1/n}$$

$\alpha^*=1-(1-\alpha)^k$: site-wide significance level, the probability of at least one of *k* comparisons being significant

CNWRA
A center of excellence in earth sciences and engineering®

Verification Resampling

- Without resampling

$$\alpha^* = 1 - (1 - \alpha)^k$$

- With a single verification sample

$$\alpha^* = 1 - [1 - \alpha + \alpha(1 - \alpha)]^k$$

- With two verification samples and one exceedance

$$\alpha^* = 1 - [1 - \alpha + \alpha(1 - \alpha)^2]^k$$

CNWRA
A center of excellence in earth sciences and engineering®

Reference

- Isaaks, E.A. and R.M. Srivastava. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY. 1990.
- Montgomery, D.C., and G.C. Runger. Applied statistics and probability for engineers. Third Edition. John Wiley & Sons. New York, NY. 2002.
- EPA. Statistical analysis of groundwater monitoring data at RCRA facilities—Unified guidance. EPA–530–R–09–007. EPA Office of Resource Conservation and Recovery Program Implementation and Information Division. Washington DC. 2009.