November 2015

Exploring and Monitoring Collection Data

By; Gerald Wilgus

Objectives

- Review of Basic Statistics;
  - Parameters and Statistics
  - Uncertainty and Confidence
- The “Flow” of Data Assessment and Analysis;
  - Stage 1; Data Organization
  - Stage 2; Selection of Initial Descriptive Techniques
  - Stage 3; Predictive Measurements
    - The Watersheds as Processes
- Ongoing Work

Basics

Parameters, Statistics and Data

- Parameters are known values of a population
- Statistics are point estimates of parameters based upon a sample of a population

Value	Parameter	Statistic
Central Tendency	μ	Mean (X-Bar)
Dispersion	σ	Standard Deviation (s)

- Data can as Variables (continuous) or Count (non-negative integers)

Variability and Uncertainty

- Variability as a measure of dispersion
  - Our components of variability
    - Sample site conditions
    - Sampling from a population
    - Assay/Identification
- Variability provides uncertainty in our estimates
  - The estimated amount by which an observed or calculated value may differ from its true value
  - s is not an unbiased estimate of σ

$$\hat{\sigma} = \left[1 + \frac{1}{4(N-1)}\right] \cdot s$$

Confidence

- Confidence is Related to Error

		Reality	
		Actually True	**Actually False**
Decision	**Accept as True**	Correct	β error
	Reject as False	α error	Correct

- Confidence:
  - In making the right decision to accept something = 1 - alpha
  - In making the right decision to reject something = 1 - beta

So; What Do We Need To Know?

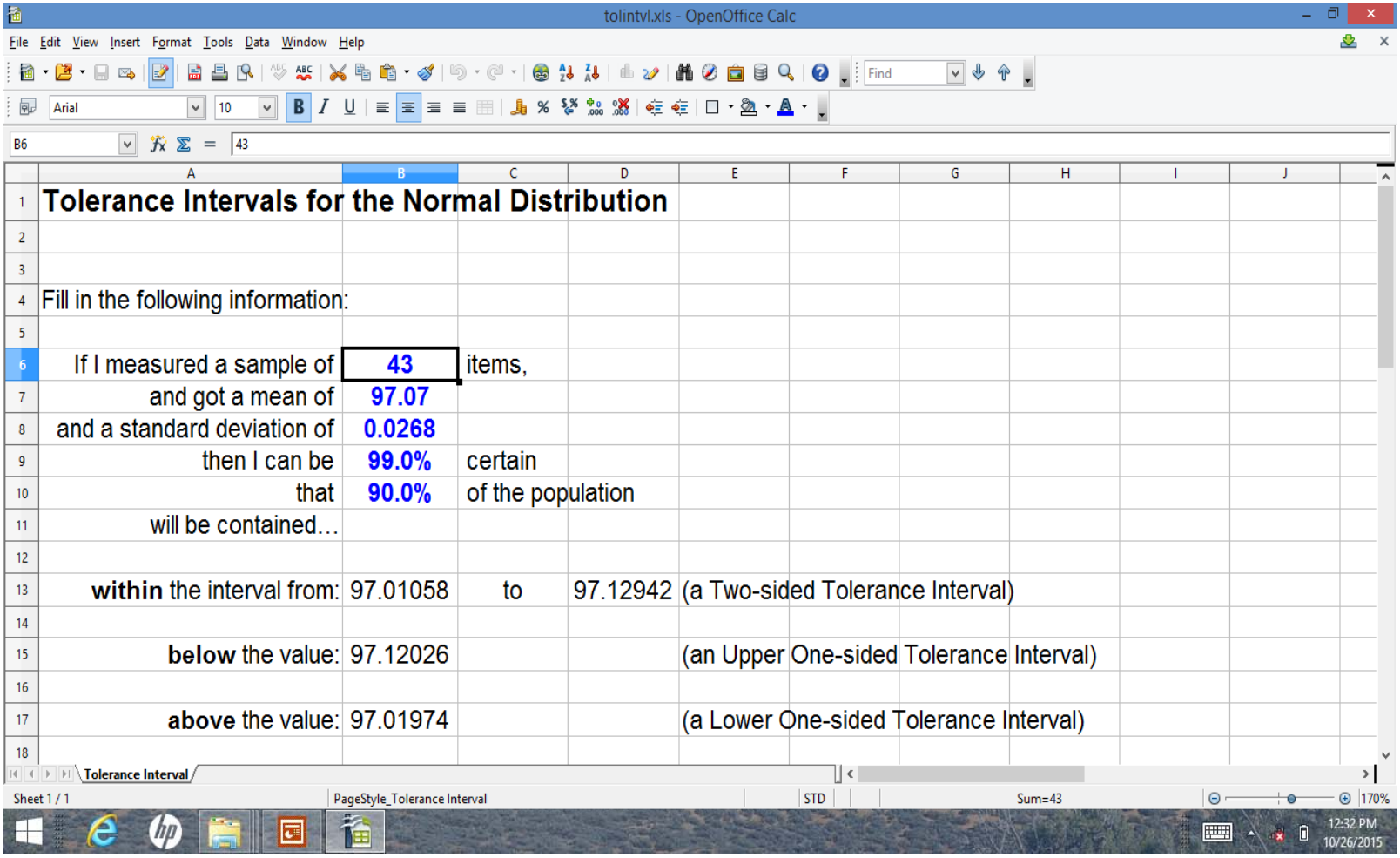
- Control sources of variability in our collection to enhance **Reproducibility and Repeatability.**
- Understanding that we are working with semi quantitative count data.
- Recognize uncertainty and confidence in statements about our data.

Digression 1

A Common Error: Confusing Parameters and Statistics

- People sometimes will say that if they know the historical spread of data, any new sample data point is expected to be within 2 standard deviations (95%) of that spread.
  - Makes a parametric statement about population coverage.
  - Neglects Uncertainty and Confidence of the sample
- The actual statistic to use is the **Tolerance Interval**
  - If I have **n samples,** then I can be **X% certain** (confidence) that **Y% of the population** (statistic) will be contained in the **range of A to B.**
  - So, new samples that are within the range we will accept as being from the same population

Tolerance Interval Worksheet



Digression 1
What Happens in Practice?

- We have done 5 collecting events at a site (n=5) and have an average SQI of 44.1 (X-bar) with a standard deviation of 2.26 (s). After the next event we calculate a SQI of 37.4.
  - Using the mean -/+ 2s we have a range for ~95% of the population as 39.6 to 48.6
    - We would fail to accept the new datum as coming from the historical population.
  - Using a Tolerance Interval for 95% of the population at 95% confidence we have a range for the population of 32.6 to 55.6.
    - We would accept the new datum as part of the population as we currently estimate it.

Analysis of Our Data

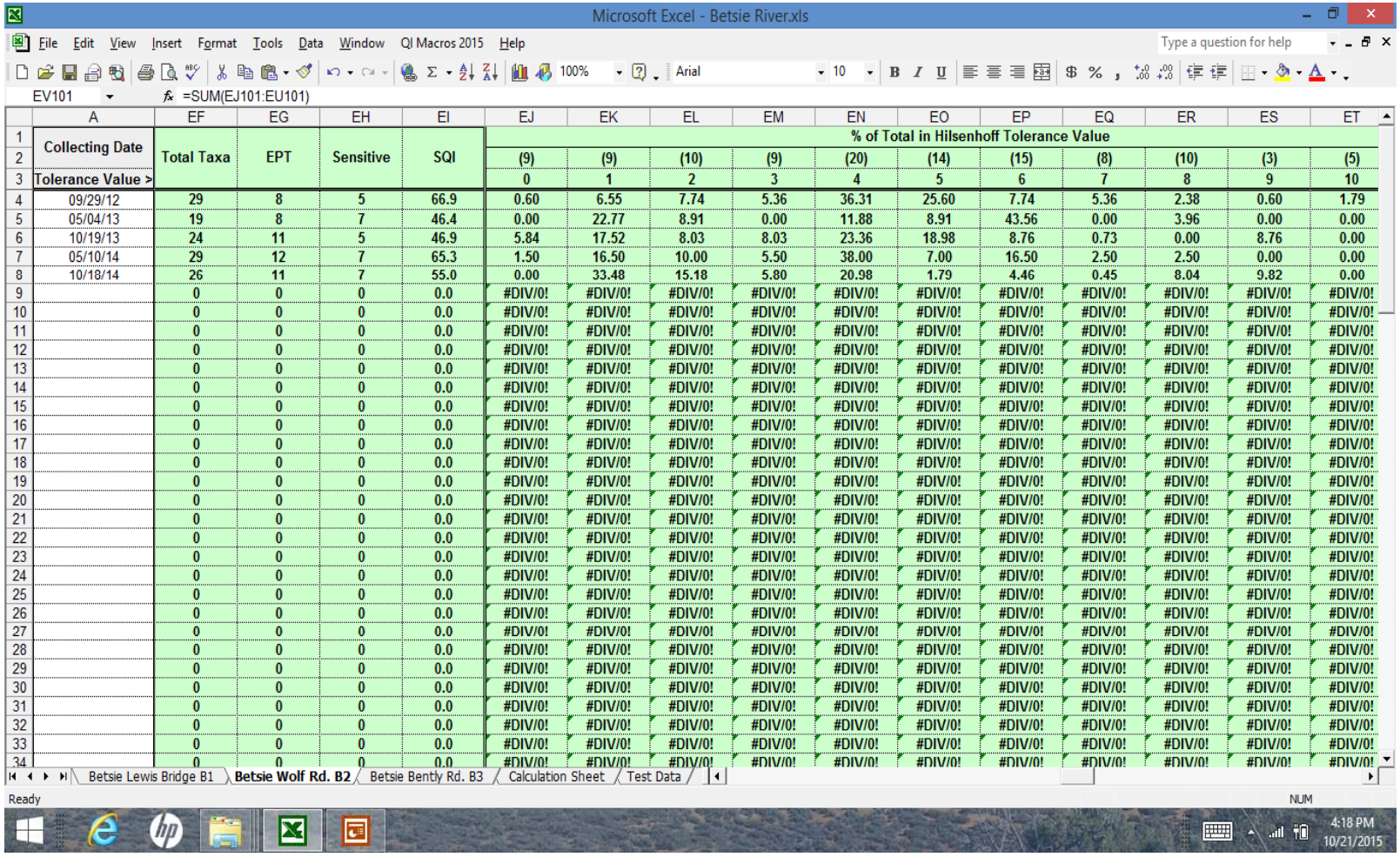
Data do not give up their secrets easily. They must be tortured to confess.

Jeff Hopper, Bell Labs

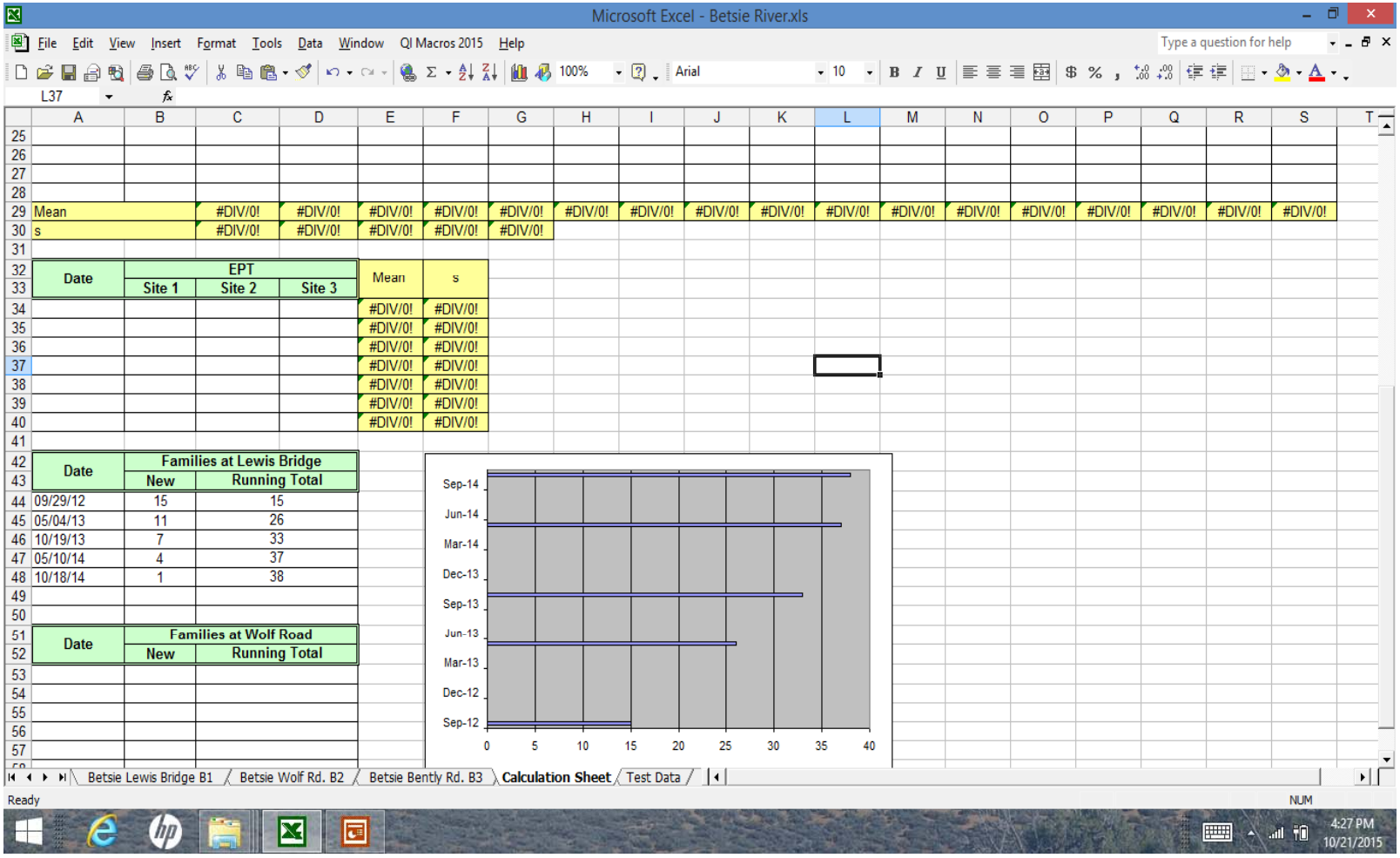
Develop The Analysis Systematically

- Organize The Data
- Define the question
- Understand the assumptions
- Analysis
  - Explore the data
  - Conduct analysis
    - Beware of confirmation bias
    - Was the objective of the question answered?
  - What are the conclusions?
- Target the presentation to the audience
- Check with somebody with knowledge and experience in applied statistics.

Set Up For Data Analysis



A Calculation Scratch Pad is Handy



Tools for Understanding our Data

- Graphical
  - Bar Charts
  - Radar Charts
  - Box and whisker plots
  - Run charts or other operational plots
- Analytical
  - Tests
    - Tolerance Interval
    - Hypothesis Testing
  - Relationships
    - Correlation

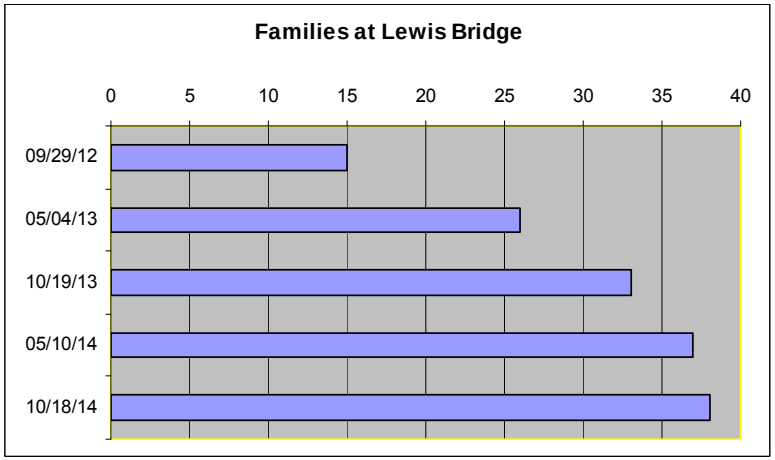
Graphical Analysis

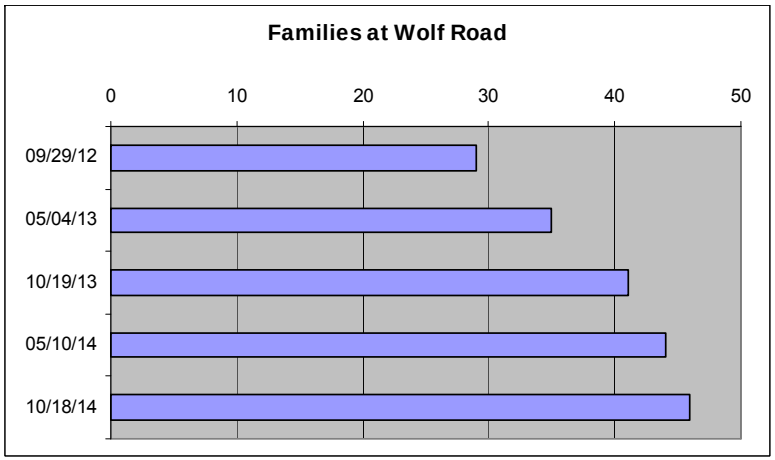
- Visual demonstration of relationships
- Can provide clues for further investigation
- Targets an audience which may not have sophistication in data analysis, statistical or otherwise
- Caution!
  - The manner of graphic presentations can bias interpretation:
    - Prominence of left to right, or top to bottom presentation
    - Scaling
    - Hiding relationships within noise.

Example: Testing an Assumption

- If we want to ask whether or not the distribution of families or Hilsenhoff tolerance values will provide the "fingerprint" of a site, we assume that we account for a knowledge of families we find at a site.
- Does a single collection event yield a view of all the families one finds at a site?
  - We are fighting variability
    - Niche sampling
    - Seasonality variability
    - Limitations of collectors/pickers
  - YES; predict few new families at a site over time
  - NO: predict accumulation of families over time

We See Accumulation of Families That Slows Over Time

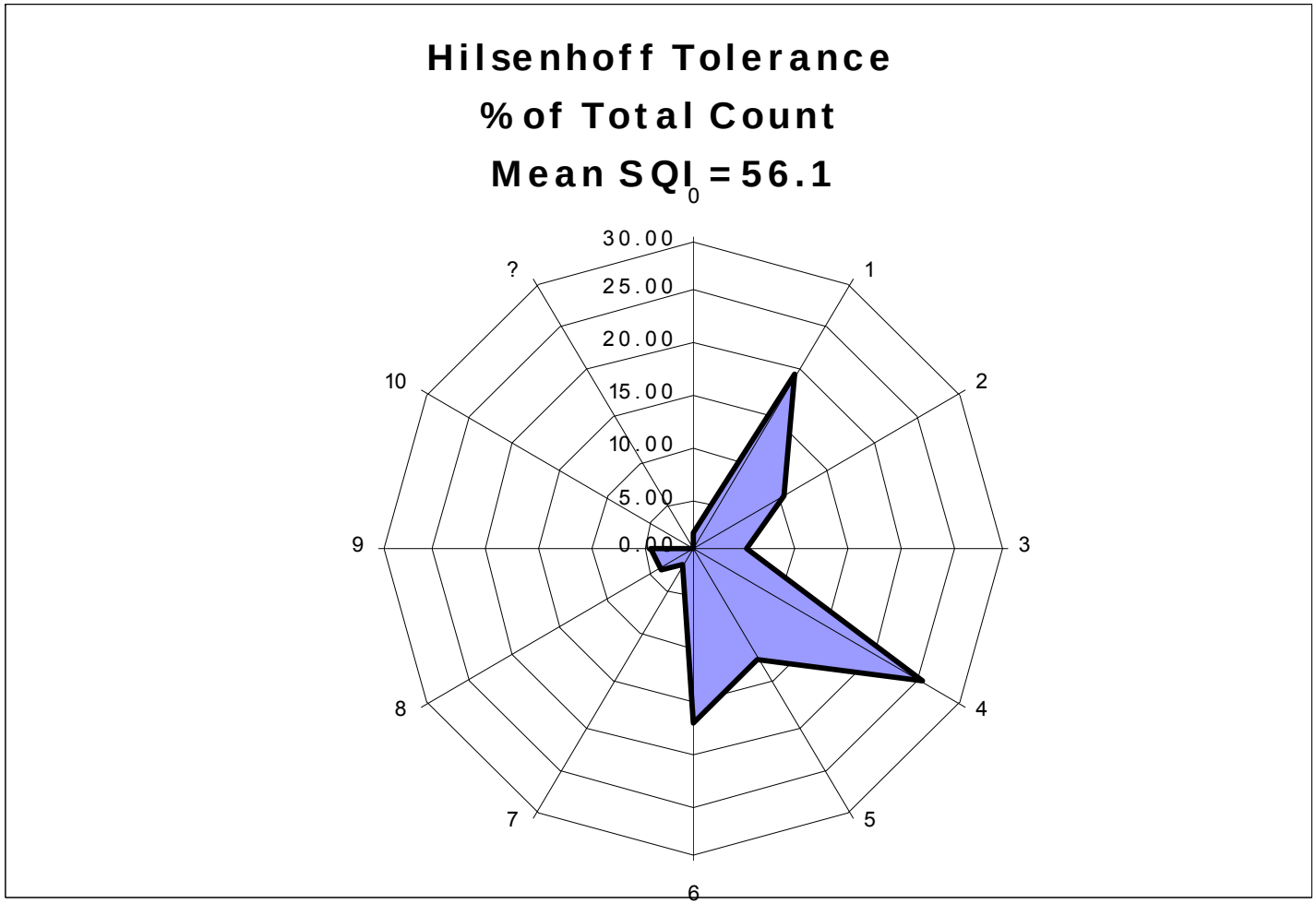




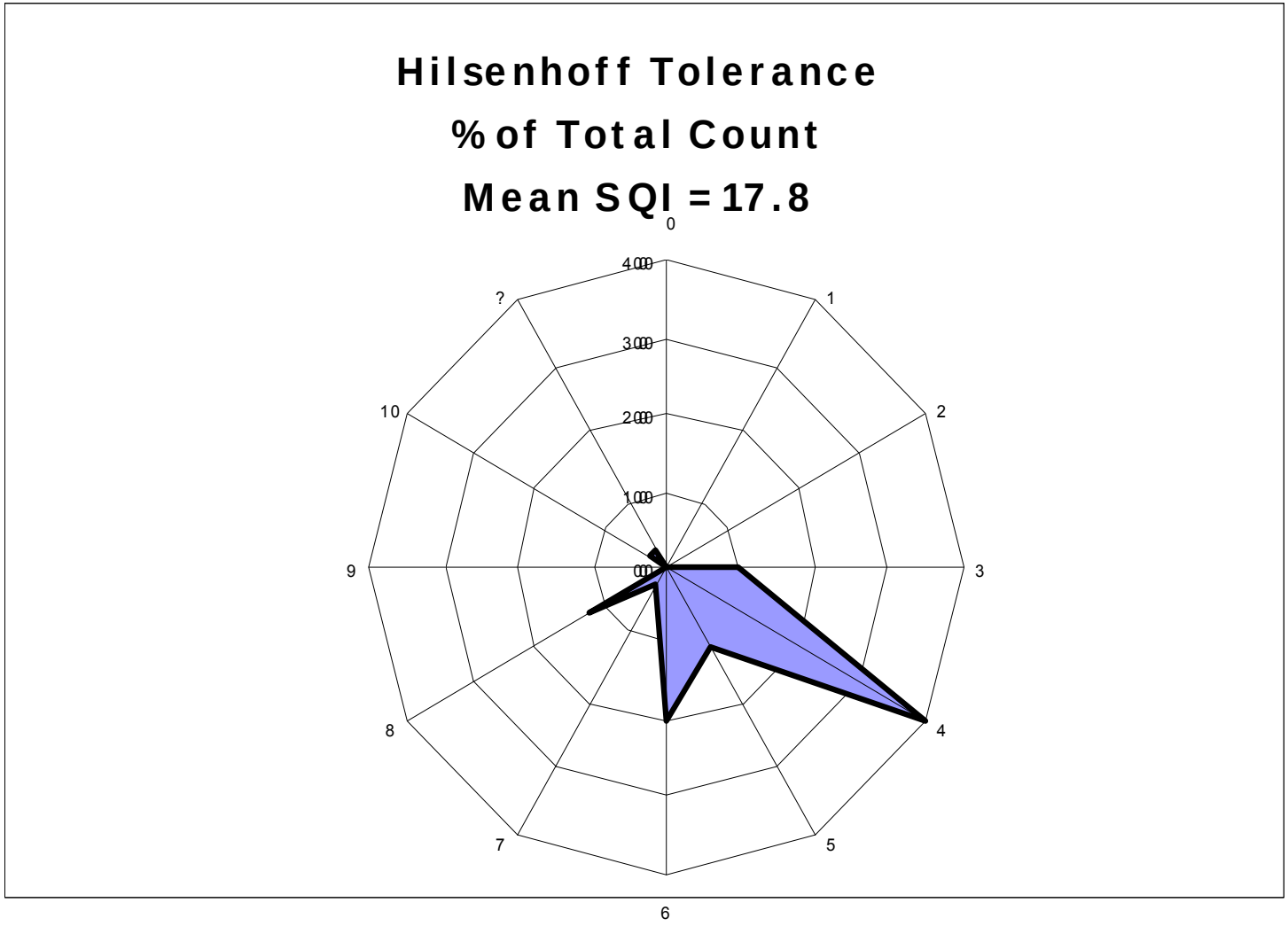
What Does This Tell Us?

- Long term collection evenutally details all families at a site.
- Perhaps we can compare relative distribution of families as fingerprints of a site.
- Can a correlation be graphically presented?
  - We assume that the tolerance of orders/families relate to our SQI.
  - What graphic to use in investigating this; especially when we have yet to apply quantitative hypothesis testing?
    - Bar charts present interpretive biases.
    - Need to enhance differences.
    - A Radar Chart that enhances proportional differences. [A simple Excel chart.]

Radar Chart for Comparison; Betsie Site 2



Radar Chart for Comparison; Extracted Data, Site X

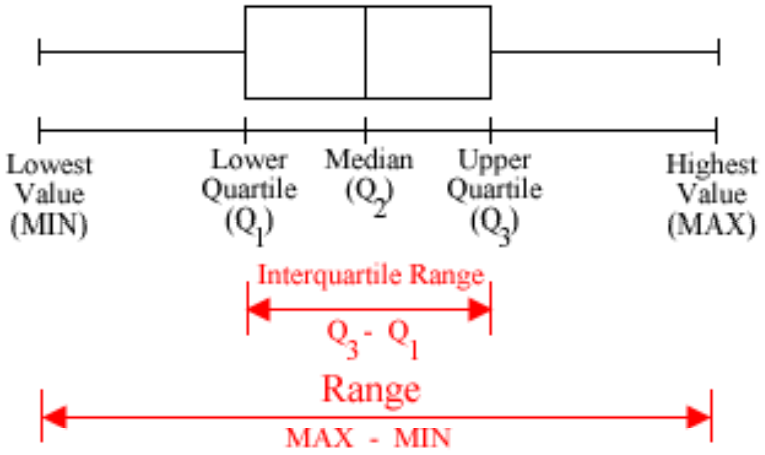


Beyond Description: Detecting Difference

- Box Plots are uniquely adapted to create a visual reference for detecting differences and opportunities for investigation.
- Focuses on the central tendency of the data.

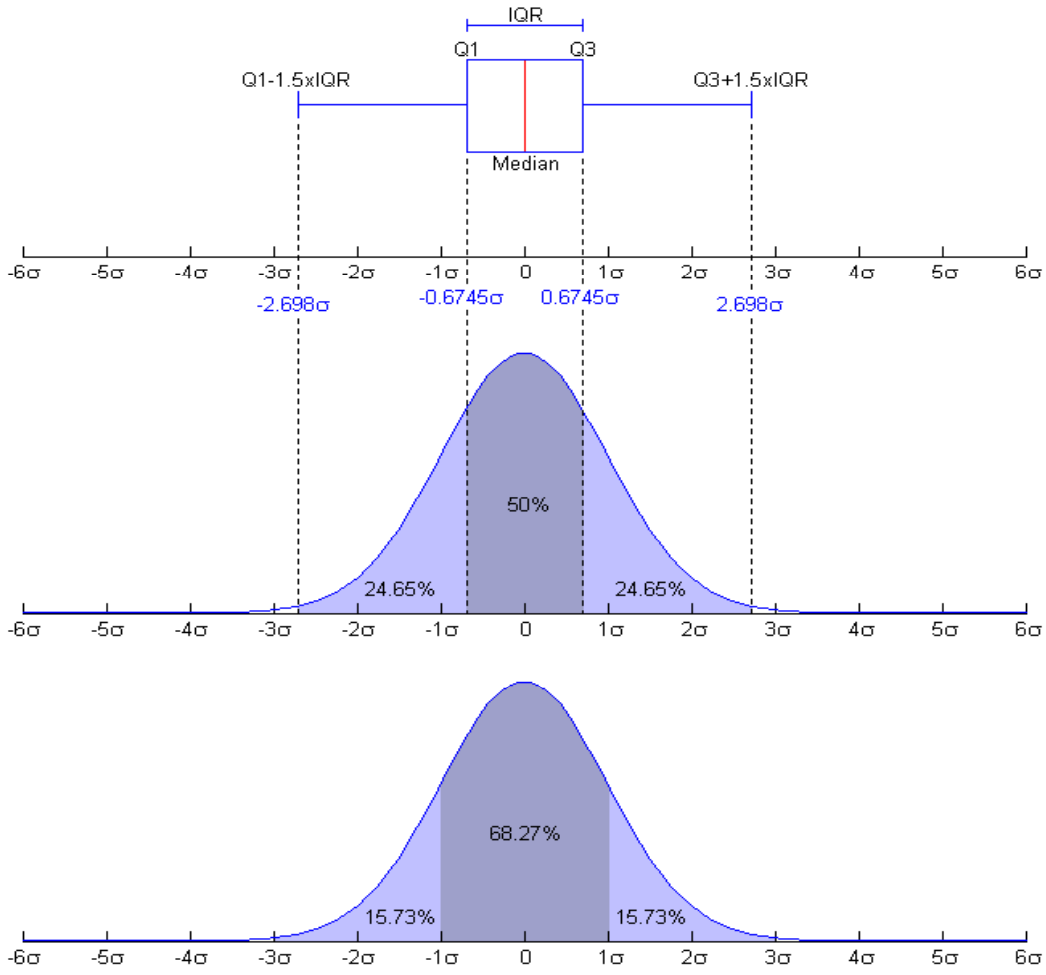
Boxplot

- Provides Information about
  - Dispersion
  - Skewness
  - Outliers
- Useful for
  - Quick visualization
  - Comparisons of data

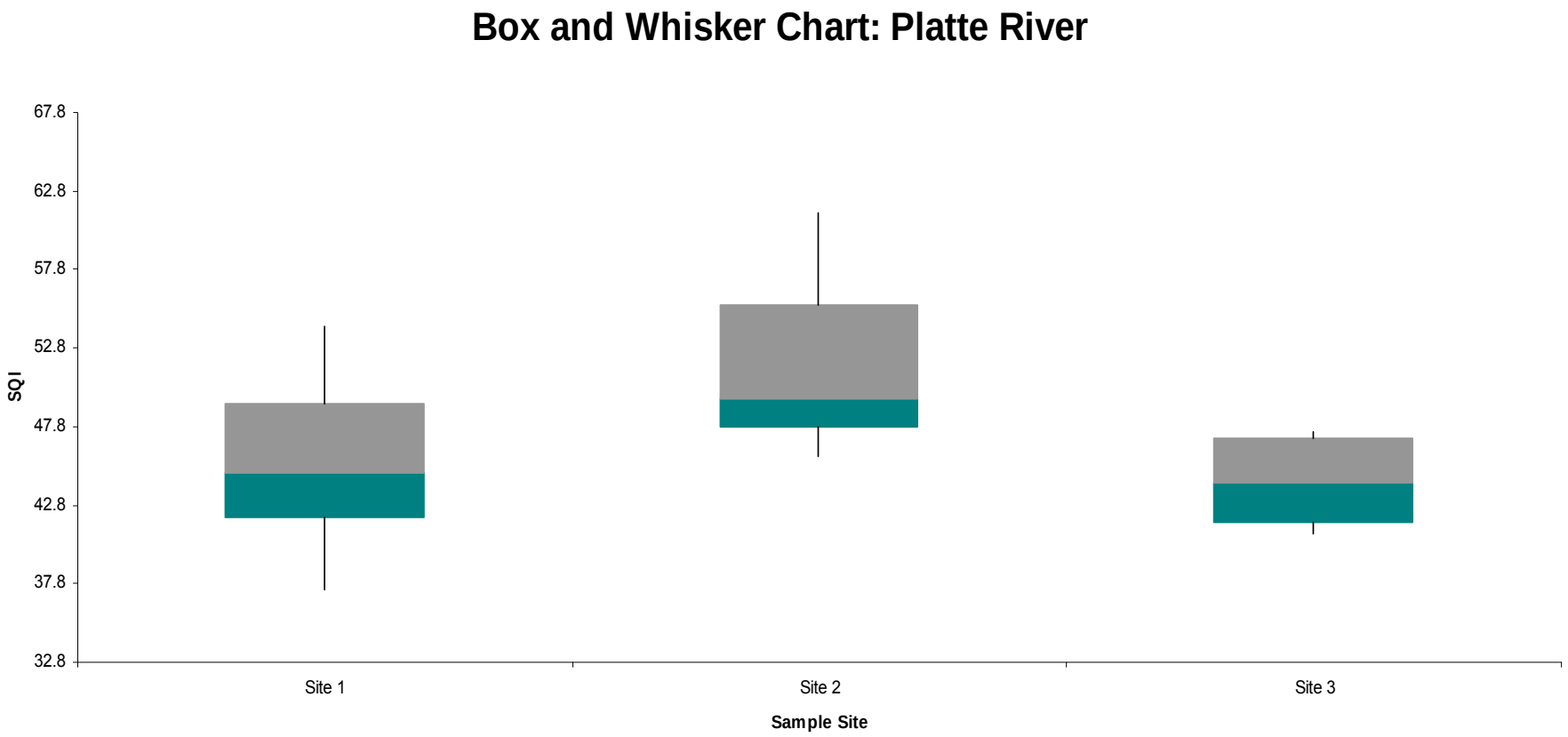


Boxplot and Histogram

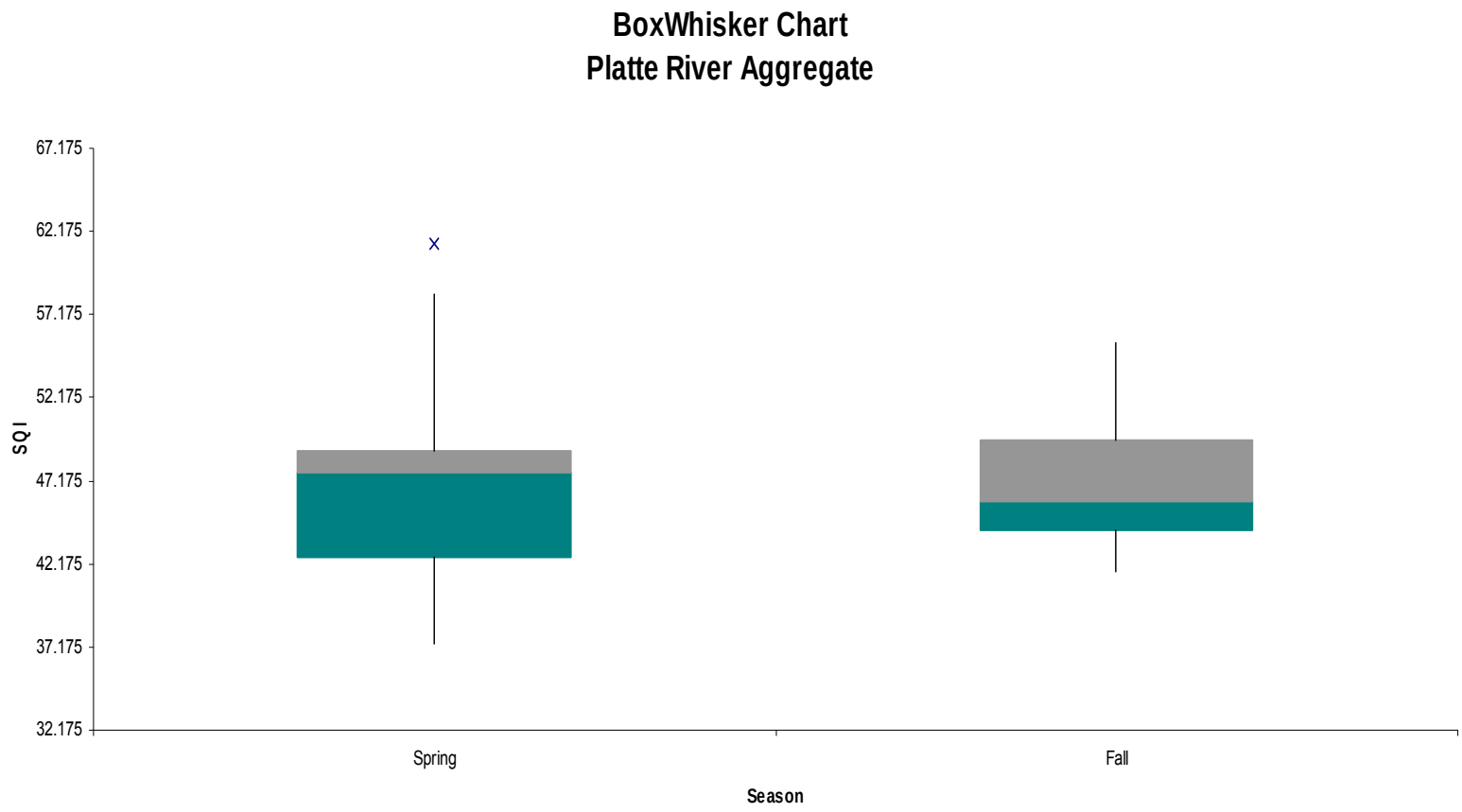
(From Magnus Manske)



Comparison of Site SQI: Platte River

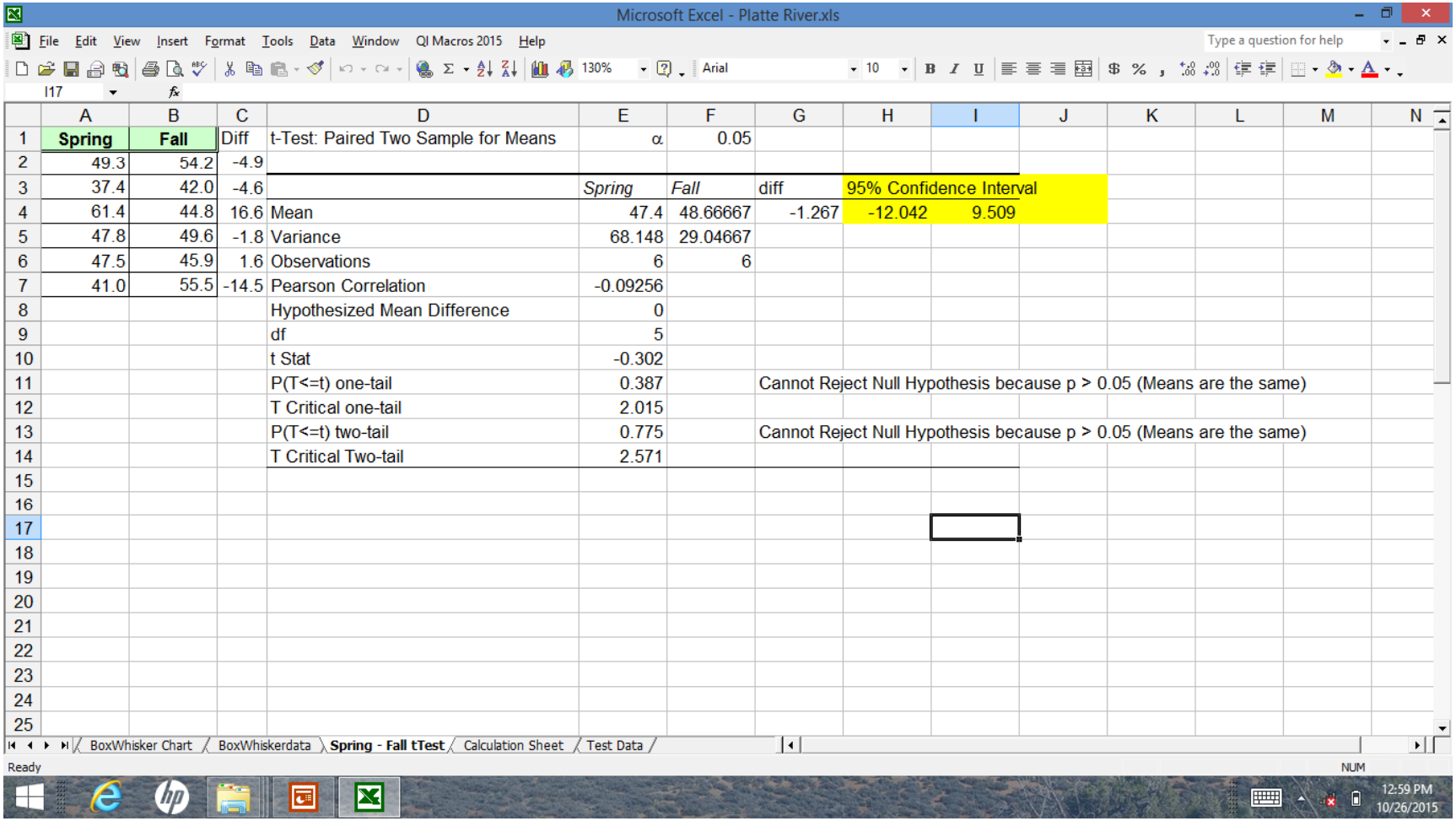


Seasonal SQI: All Sites



Does This Check Out?

- A sensitive analytical test is the paired comparison t-test.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Spring**	**Fall**	Diff	t-Test: Paired Two Sample for Means	α	0.05								
2	49.3	54.2	-4.9											
3	37.4	42.0	-4.6		*Spring*	*Fall*	diff	95% Confidence Interval						
4	61.4	44.8	16.6	Mean	47.4	48.66667	-1.267	-12.042	9.509					
5	47.8	49.6	-1.8	Variance	68.148	29.04667								
6	47.5	45.9	1.6	Observations	6	6								
7	41.0	55.5	-14.5	Pearson Correlation	-0.09256									
8				Hypothesized Mean Difference	0									
9				df	5									
10				t Stat	-0.302									
11				P(T<=t) one-tail	0.387		Cannot Reject Null Hypothesis because p > 0.05 (Means are the same)							
12				T Critical one-tail	2.015									
13				P(T<=t) two-tail	0.775		Cannot Reject Null Hypothesis because p > 0.05 (Means are the same)							
14				T Critical Two-tail	2.571									
15														
16														
17														
18														
19														
20														
21														
22														
23														
24														
25														

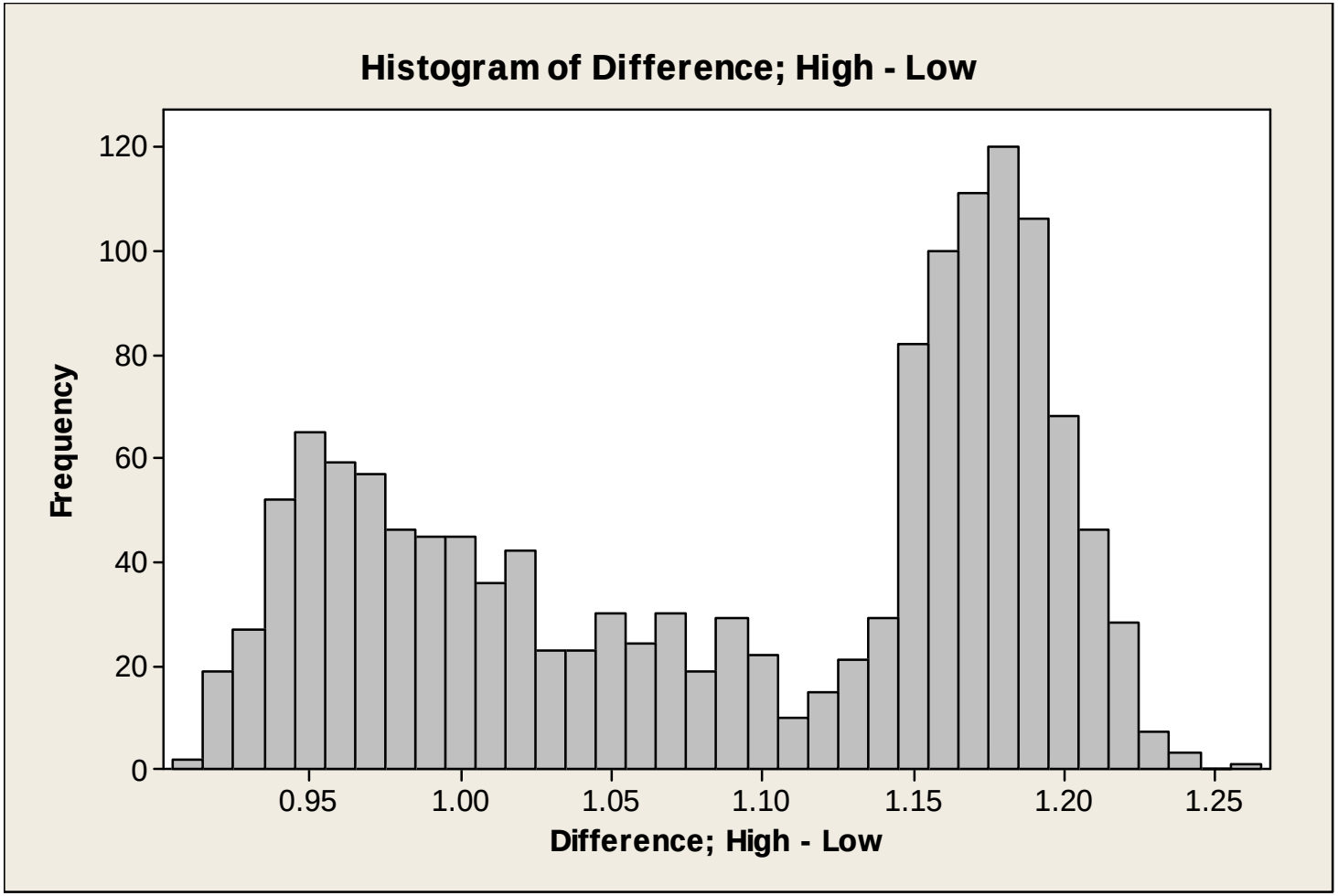
Detecting Change

Check Each New Datum Against History

- Does it fit into a reasonable tolerance interval for the SQI of a site?
- Does it fit within the interquartile range or the full distribution of a Box and Whisker plot of the SQI for a site?

- Caution!
  - Results must be considered provisional until over 20 data points are available.
  - Do not overreact; investigate likely causes of variation first.
    - Issues with sampling Reproducibility and Repeatability.
    - Look at the historical distribution of values.

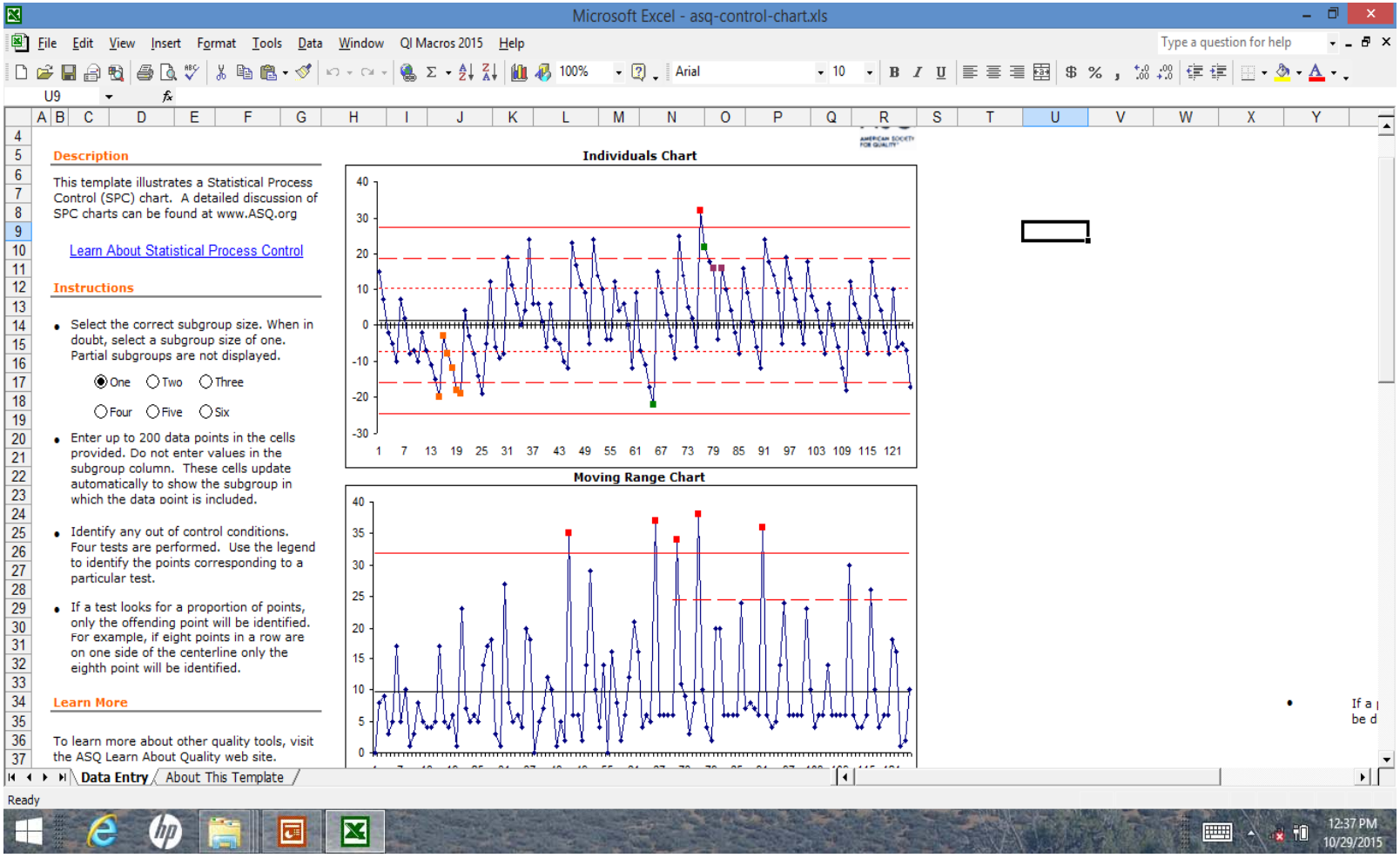
Digression 2: The Histogram



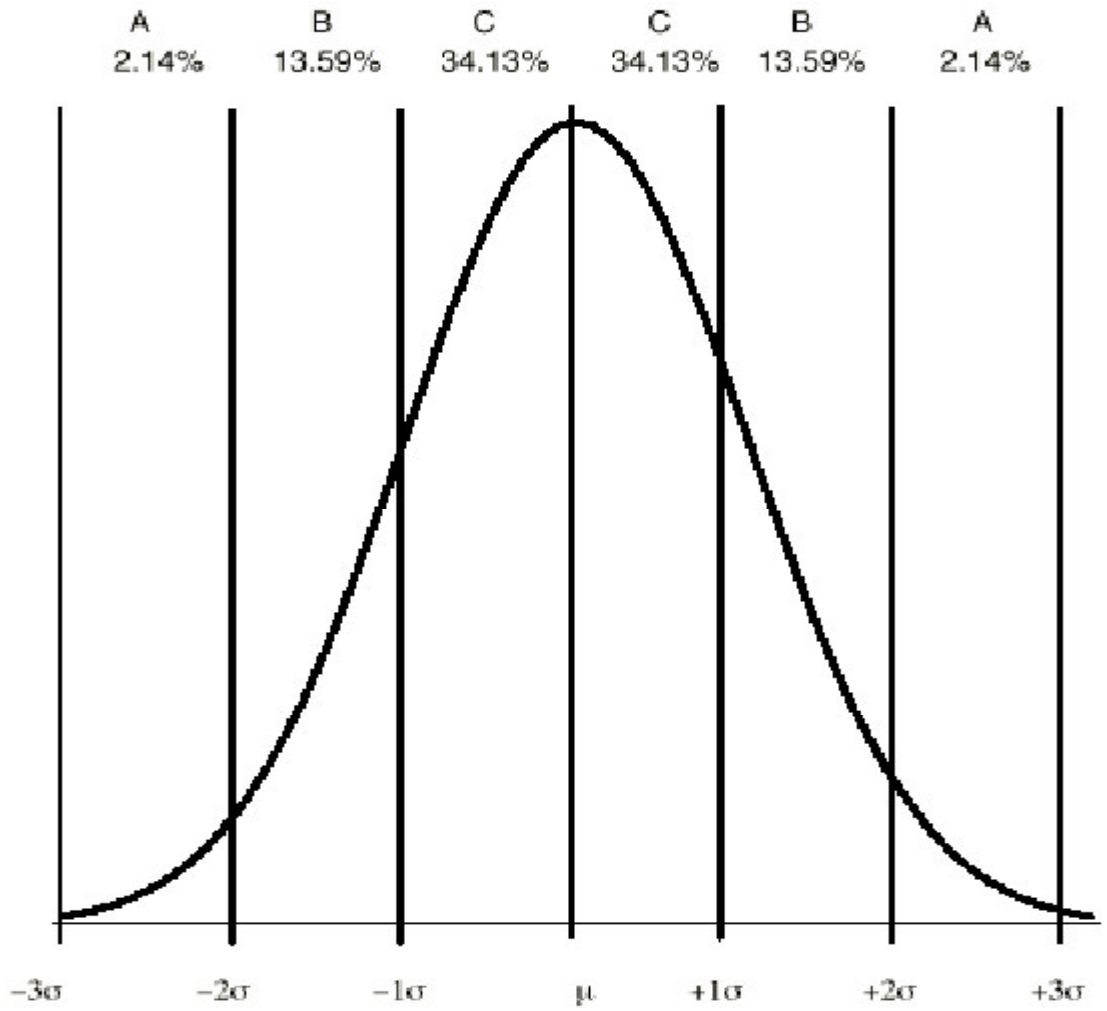
Control Charting

- We can treat our streams as processes that result in the selection of species adapted to its conditions.
- Control charting allows us to distinguish between normal variation and special causes.
  - Many types of control charts are available to use.
  - Simple to set up.
  - Requires experience to use them in making a conclusion that change has happened. The simplest are susceptible to overreaction.
  - All control charts must be treated as provisional until over 20 data points are obtained.

A Simple Template From the ASQ



How Do Control Charts Work?



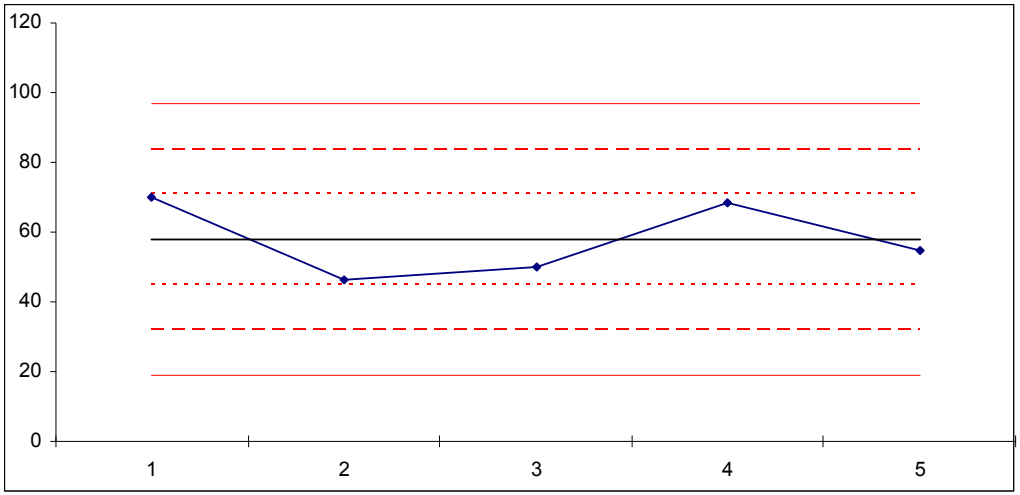
Control Chart Decision Rules

- A stable SQI will show:
  - Most points are near the average.
  - A few points are near the control limits.
  - No points are beyond the control limits.
- Simple Change is detected when:
  - One point exceeds the control limits.
  - Nine or more points are on the same side of the mean.
  - Two out of three points are more than 2 s of the mean in the same direction.
  - For more rules, see Nelson Rules; https://en.wikipedia.org/wiki/Nelson_rules
- **Apply With Caution**

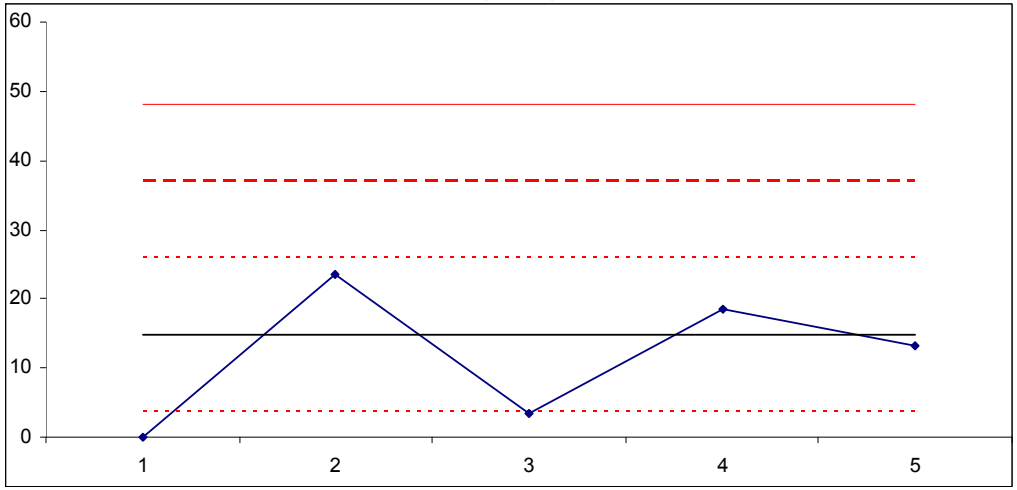
Setting Up The Control Chart

- What organization makes sense? Individual data points or subgroups
  - Individuals. **I,MR Control Chart**
  - Yearly subgroups. **X-bar, R Control Chart**
- Using special control charts
  - Detect small changes? **MA Control Chart**
  - Emphasize the current datum? **EWMA Control Chart**
  - Sensitive to correlations? **Hotelling T2 Control Chart**

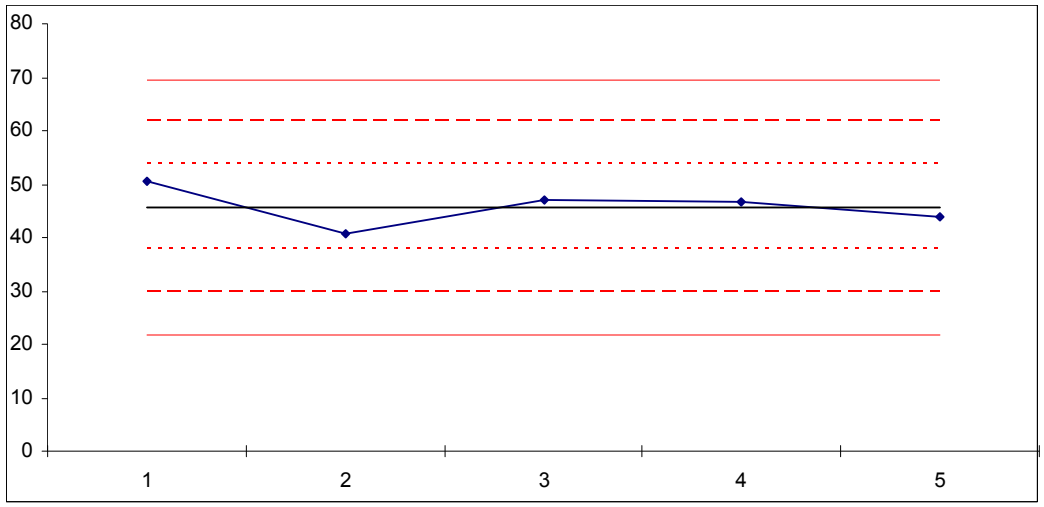
A Simple Chart: Individual, Moving Range SQI Betsie River, Site 2



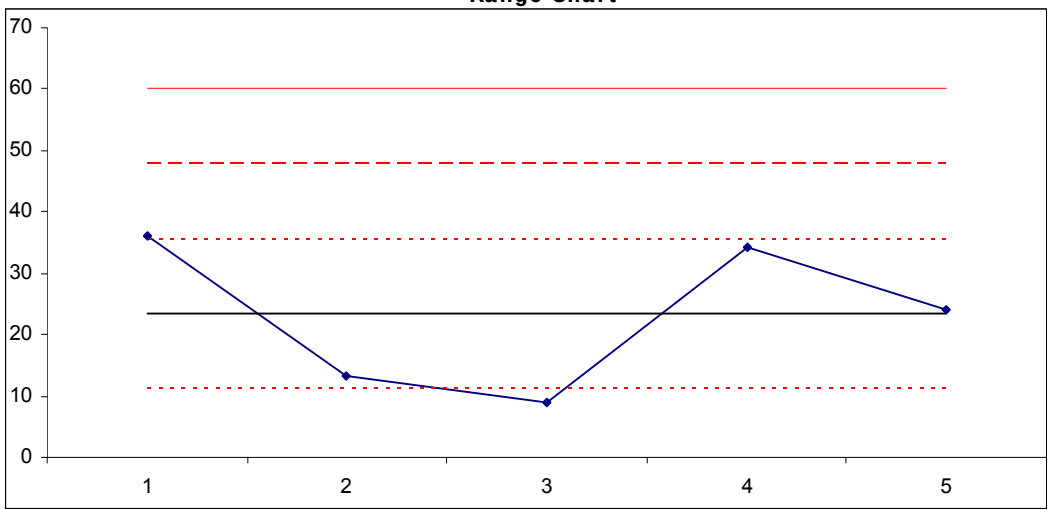
Moving Range Chart



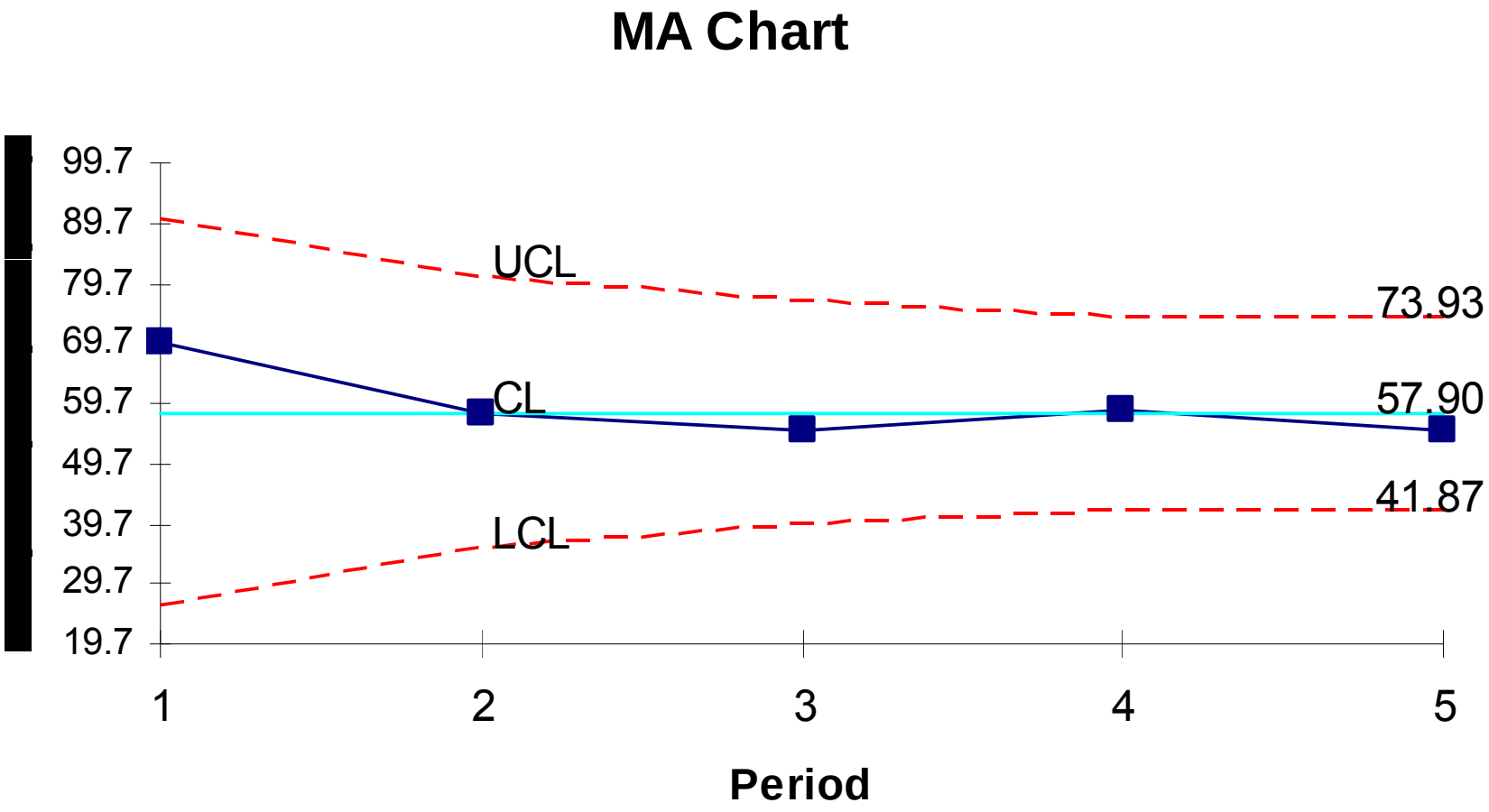
X-bar R Chart
Betsie River [Example Only]



Range Chart



The Moving Average
Betsie River, Site 2



Future Work

- Create a common workflow for analysis (publish for MiCorps users)
- Correcting for different numbers of families within each Hilsenhoff Tolerance Value to improve site fingerprinting.
- Identification of factors that correlate with SQI.
- Investigation of control charts sensitive to change.
  - Bivariate control charts (Hotelling T2).
  - Exponentially weighted moving average control charts.
- Would like to see your family count data and work to develop a common workflow. wilgusg@gmail.com

References

- NIST / Sematech Handbook of Engineering Statistics
  - http://www.itl.nist.gov/div898/handbook/
- Any introduction to statistical analysis;
  - *Introduction to Statistical Analysis,* Dixon, Massey, McGraw-Hill Inc.
  - *Quality Engineering Statistics,* Dovich, ASQ Press
- Statistical Quality Control Handbook (AT&T); Western Electric *(Yes – this is the granddaddy of all statistical quality references)*
- The Desk Reference of Statistical Quality Methods; Crossley, ASQ Press

Questions ???

Thank You