

Institute for Tribal Environmental Professionals
Tribal Air Monitoring Support Center
Melinda Ronca-Battista

Brenda Sakizzie Jarrell
Southern Ute Indian Tribe

Data Smack Down **(Exploratory Data Analysis)**

Why ITEP does this:

- Assisting tribes all over country
- Similar questions are asked
- All evaluations begin the same
- Need for free (MS Office) tools
- Applicable to any data
- Supplements and uses existing EPA tools, helping tribes use AMTIC and have confidence of the interpretation of their data

All materials at:

Address



<http://www4.nau.edu/itep/resources/>



ITEP

institute for tribal environmental professionals



RESOURCES

[Data Analysis](#) [ECI](#) [ENV-Tech](#) [IAQ](#) [Maps](#) [PA](#) [Publications](#)

Home » Resources

ITEP Resources:

The resource center is a clearinghouse and point of contact for information environmental staff. We have hard copy resources available as well as the access through this page. We love to hear from you, please email, call, write recommendations, or just to say hello!

AAA Data Analysis and Interpretation folders

7-steps to data domination:

1. Clean up your data.
2. Verify QC limits met
3. Aggregate data into sets
4. Find Patterns
5. Ask your question
6. Evaluate the shapes of the distributions
7. Apply the test

Step 1: Clean Your Data

- 1. Initial Cleaning (checking links, hidden values, finding repeated rows, tracking data so none is lost)
- 2. Normalization (separating information into separate fields, using data validation to limit entries to drop-down lists)
- 3. Documenting Your Clean Data
- 4. General cleaning (macros, values-only, documentation, eliminating hidden characters)

<http://www4.nau.edu/itep/resources/>

Data Analysis, Step 1–Data Clean Up folder

Data Smack Down Step 2:QC

- Don't exert any effort on bad data—start with a quick review of QC
- Review logbooks, audits
- Generate AQS report of QC data, and
- Use EPA's DASC tool; enter data into the spreadsheet and review PLOTS

<http://www4.nau.edu/itep/resources/>

Data Analysis, Step 2-QC folder

AQS report:

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

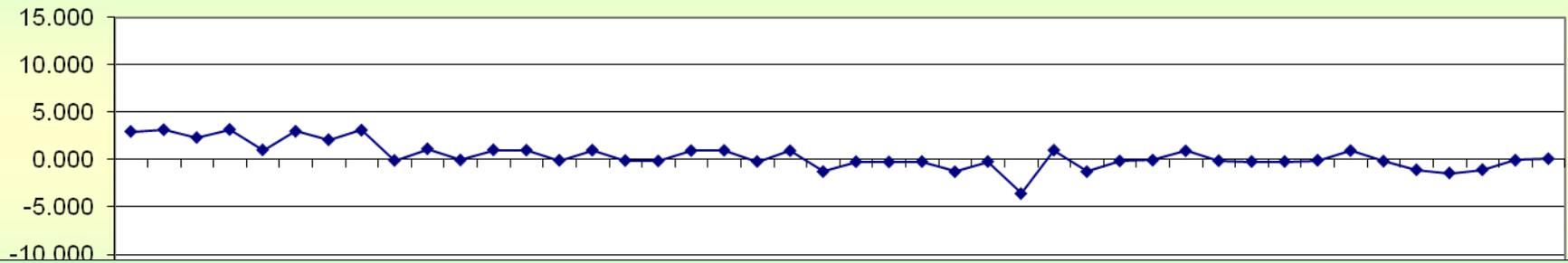
AIR QUALITY SYSTEM PRECISION REPORT

SINGLE-MONITOR PRECISION CHECKS

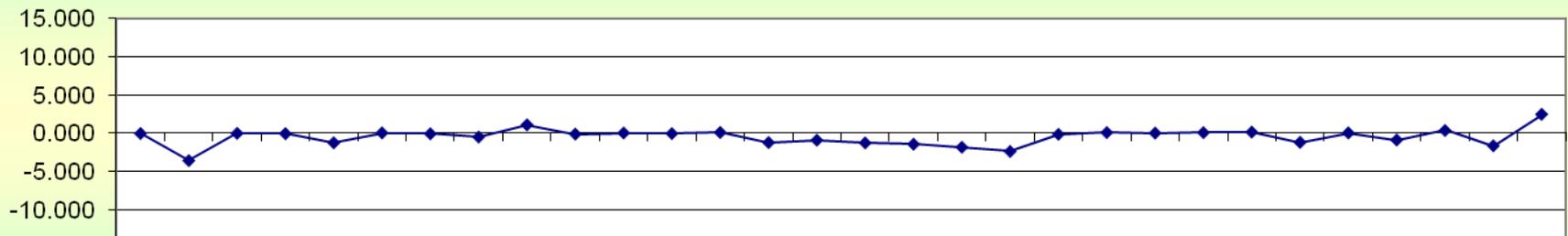
MONITOR ID	ACTUAL	MEAS.	METH CODE	%DIFF
TT-750-7001-4210:	9.00000	8.97600	054	-0.3
TT-750-7001-4210:	9.00000	9.10900	054	1.2
TT-750-7001-4210:	9.00000	9.10800	054	1.2
TT-750-7001-4210:	9.00000	9.25700	054	2.9
TT-750-7001-4210:	9.00000	9.30800	054	3.4
TT-750-7001-4210:	9.00000	9.50800	054	5.6
TT-750-7001-4210:	9.00000	9.60800	054	6.8
TT-750-7001-4210:	9.00000	8.90700	054	-1.0
TT-750-7001-4210:	9.00000	9.10700	054	1.2
TT-750-7001-4210:	9.00000	9.16000	054	1.8
				-0.5

2006:

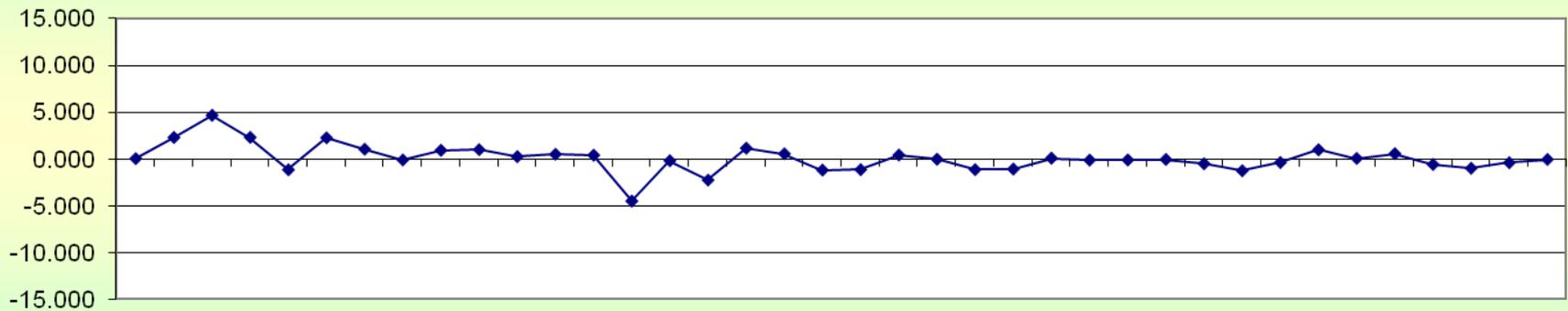
Plots generated using EPA's DASC tool:



2007:



2008:



Part of QC is completeness:

UTE 1 Ozone data						
		avg (ppb)	1st max (ppb)	4th max	observation	completeness
2005	Jan	24	45		682	91.7%
	Feb	25	46		649	96.6%
	Mar	32	58		730	98.1%
	Apr	37	63		712	98.9%
	May	36	67		739	99.3%
	Jun	32	68		713	99.0%
	Jul	33	67		732	98.4%
	Aug	26	57		739	99.3%

After Part 2-QC,



Step 3: Aggregate Data

- Hourly values slow down computer
- Use MS Access Quick Start guide to aggregate data into chunks of daily or weekly averages
- MS Access *easier than Excel* for handling missing data, excluding codes

<http://www4.nau.edu/itep/resources/>

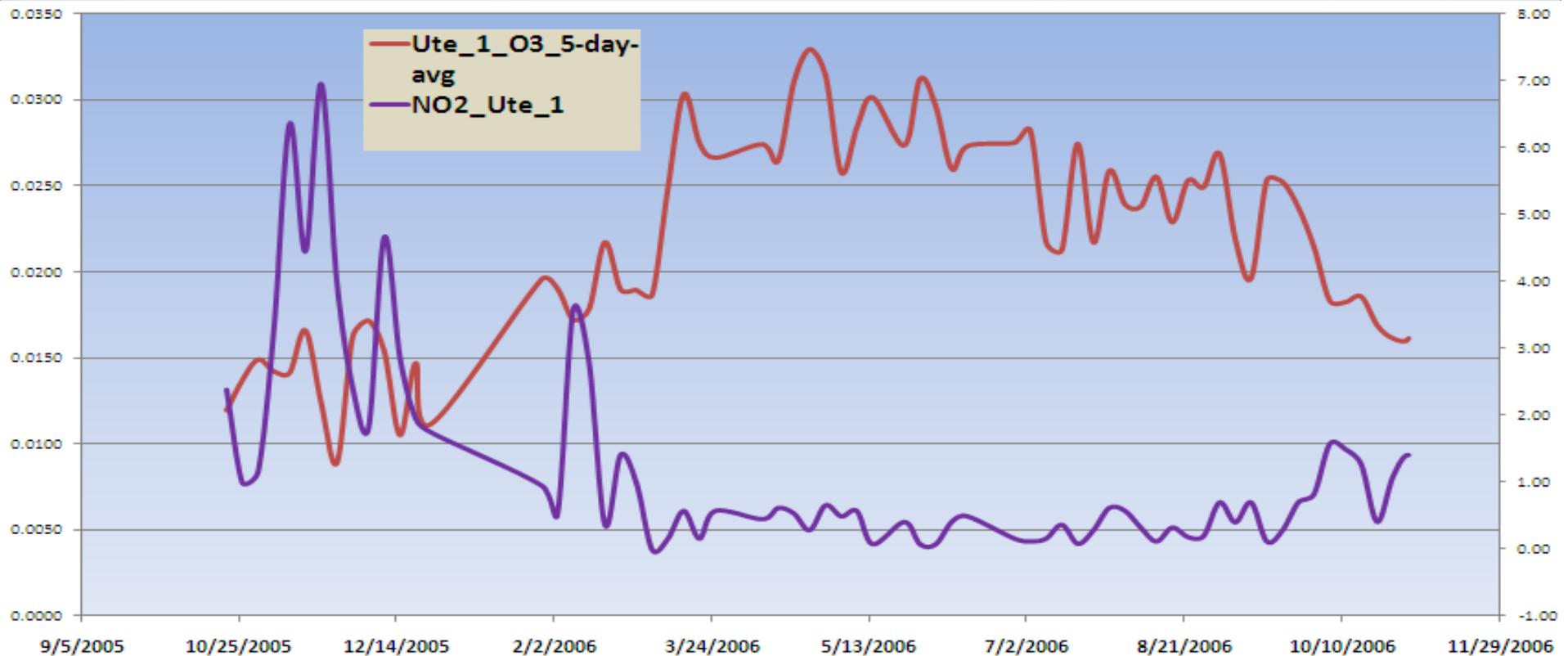
Data Analysis, Step 3–Aggregate Data folder,
Tribal Data Access Quick Start subfolder

Aggregation into:

- 5-day averages of daily max O₃ values
- Reduces number of data points from >10,000s of hourly values to ~1000s of daily values to ~100s of 5-day averages
- Enables the next step of graphing against relevant parameters (temp, solar radiation, NO₂, etc.)

Step 4–Find Patterns:

- Apply common sense to the data
- How does it vary with met parameters?
- How does it vary with other pollutants?



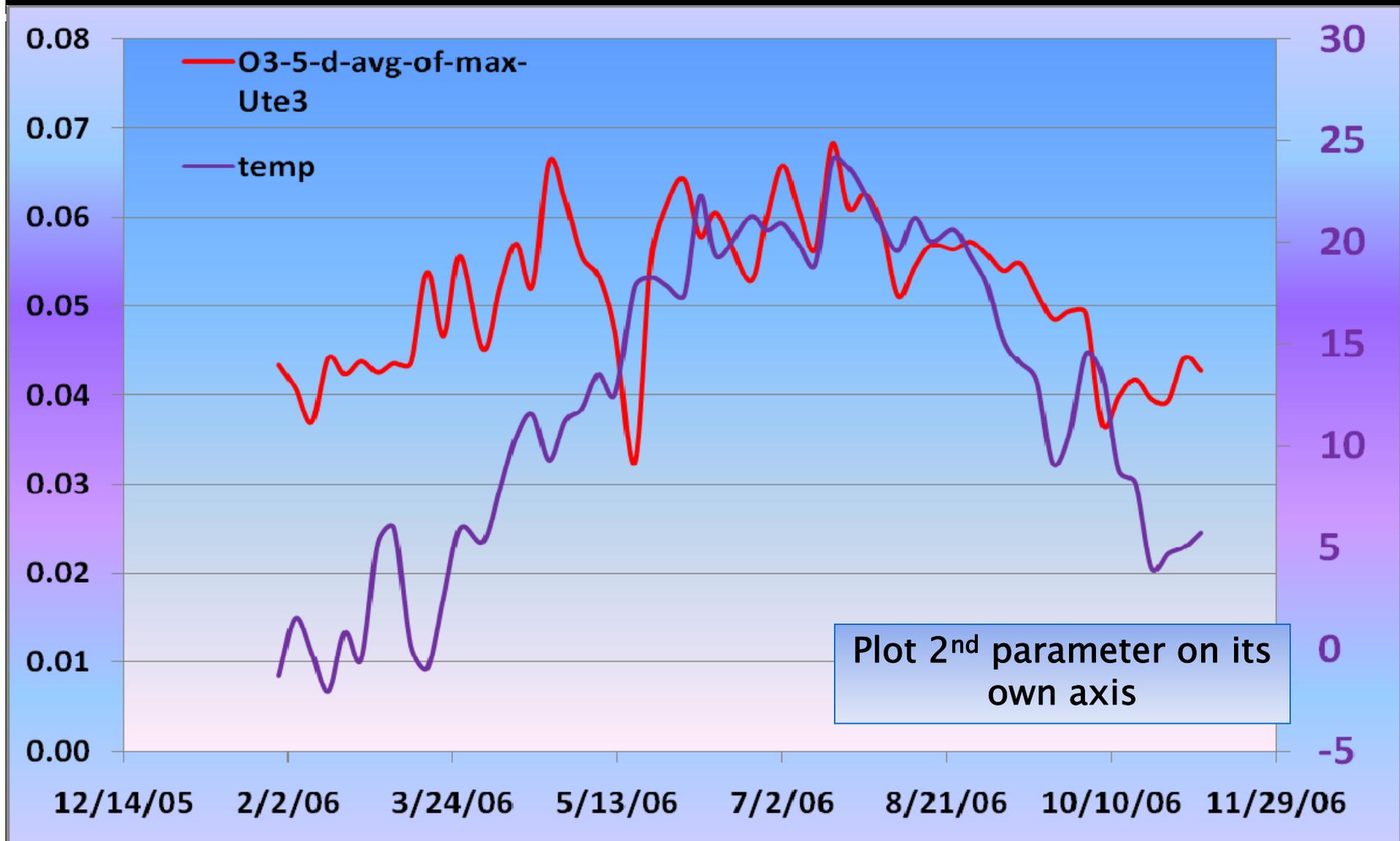
Step 4A–Use Excel Tools

- Dynamic Named Ranges
 - as your data increase, plots and summary statistics are automatically recalculated
 - See pg. 7 of doc “using ranges in excel and graphing.doc”
- Autofilter
 - Filter out or in data
 - 1–click recalculation of plots of different subsets

Step 4B–plot vs time on x–y plot:

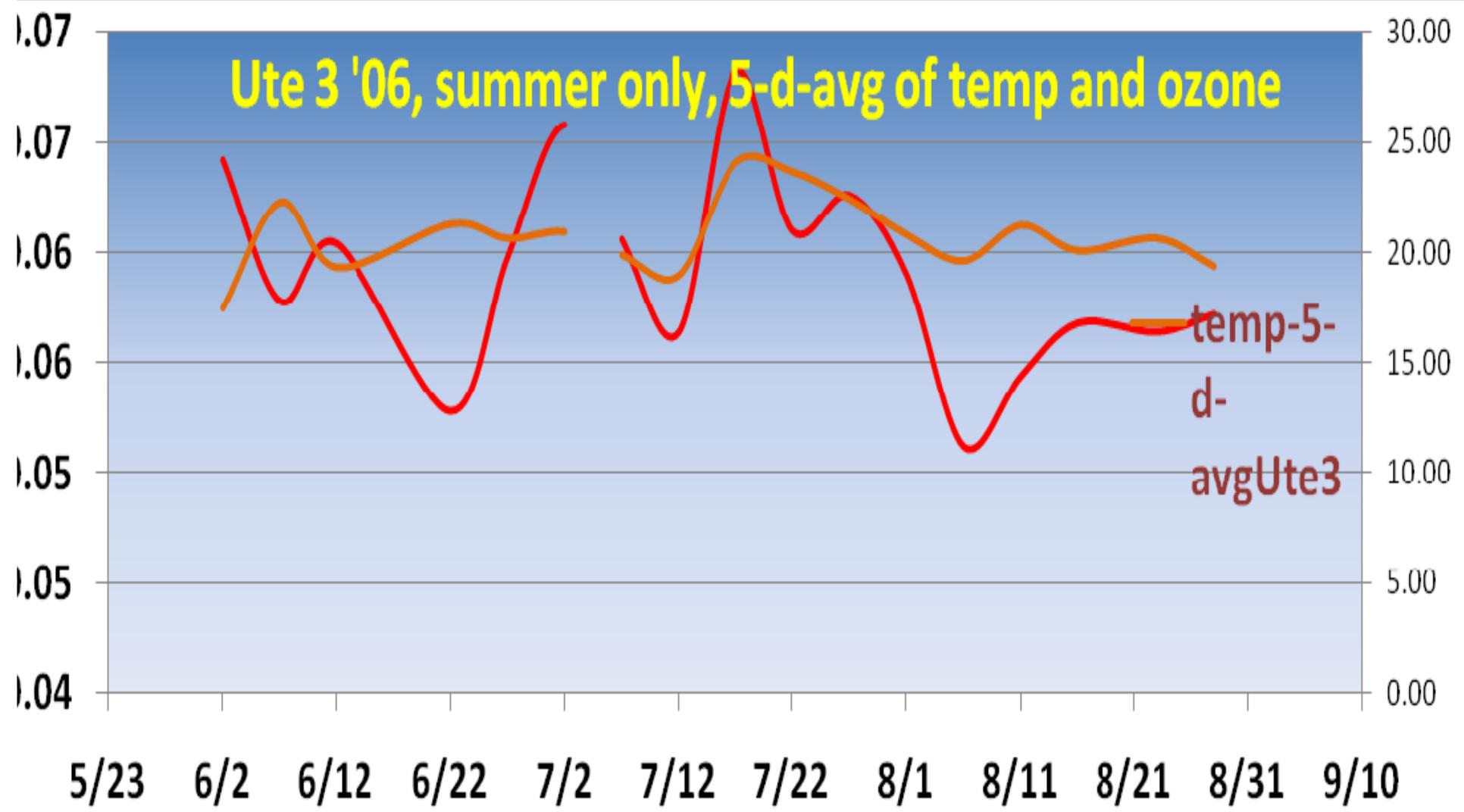
- Use x–y plot (ALWAYS SCATTERPLOT NEVER DATE because that produces category plots)
- Use secondary axis for 2nd parameter so it can have its own units
- Start with all data, then use Excel Autofilter to find subsets of data *where both parameters have values, that show a pattern*, clicking on different values that are immediately graphed

Time plot of temp and O3, 2006 only:



**With click of Autofilter
filtered/more data & look for patterns:**

ALWAYS plot on x-y
scatterplots-never dates



Step 4C—quantify the pattern

- Use linear regression between parameters
- Perfect 1-to-1 relationship with one rise on y-axis to every one run on x-axis shows:
 - Slope ~ 1 and RSQ (r^2) ~ 1
 - Can calculate in plot or using functions

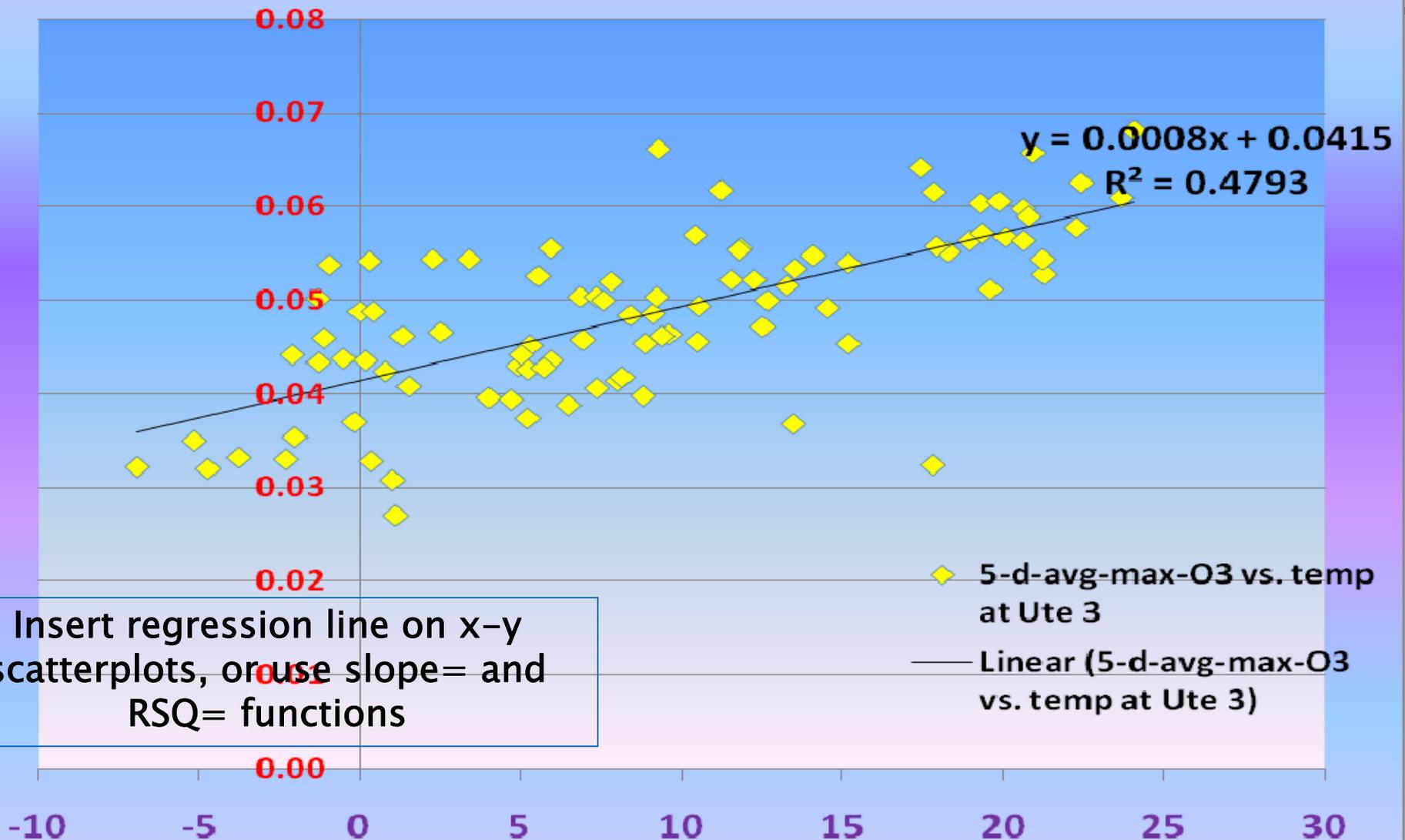
Calculate correlations:

- $=\text{slope}(Ys, Xs)$ and $=\text{RSQ}(Ys, Xs)$
- OR
- Scatterplot, show trendline, show equation and R2 on chart
- For the case of our correlation between O3 and temp, how well do they compare?



“pretty well” for all data:

5-d-avg-max-O3 vs. temp at Ute 3, all data

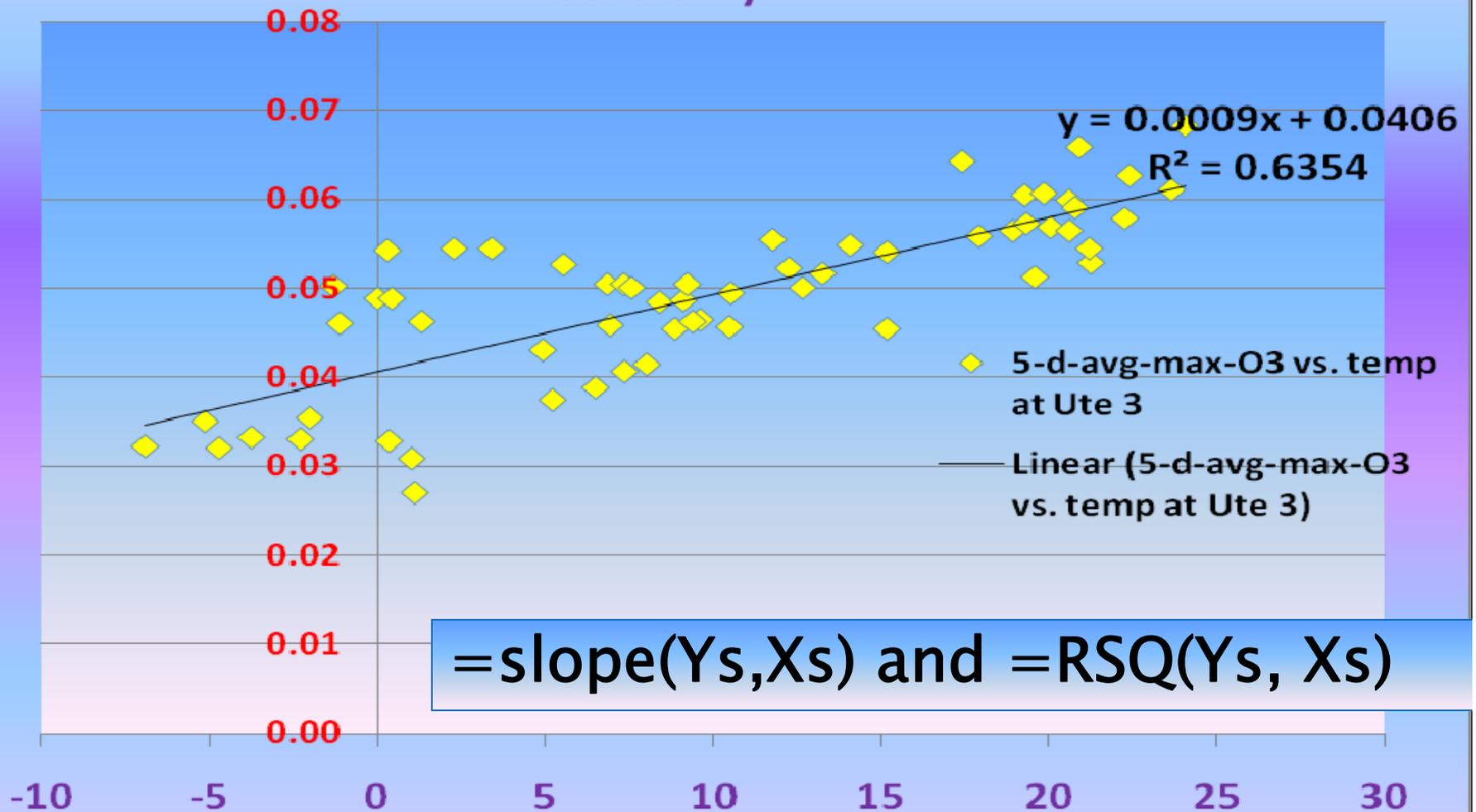


Insert regression line on x-y scatterplots, or use slope= and RSQ= functions

- ◆ 5-d-avg-max-O3 vs. temp at Ute 3
- Linear (5-d-avg-max-O3 vs. temp at Ute 3)

But with less data:

5-d-avg-max-O3 vs. temp at Ute 3, winter/summer data only



=slope(Ys,Xs) and =RSQ(Ys, Xs)

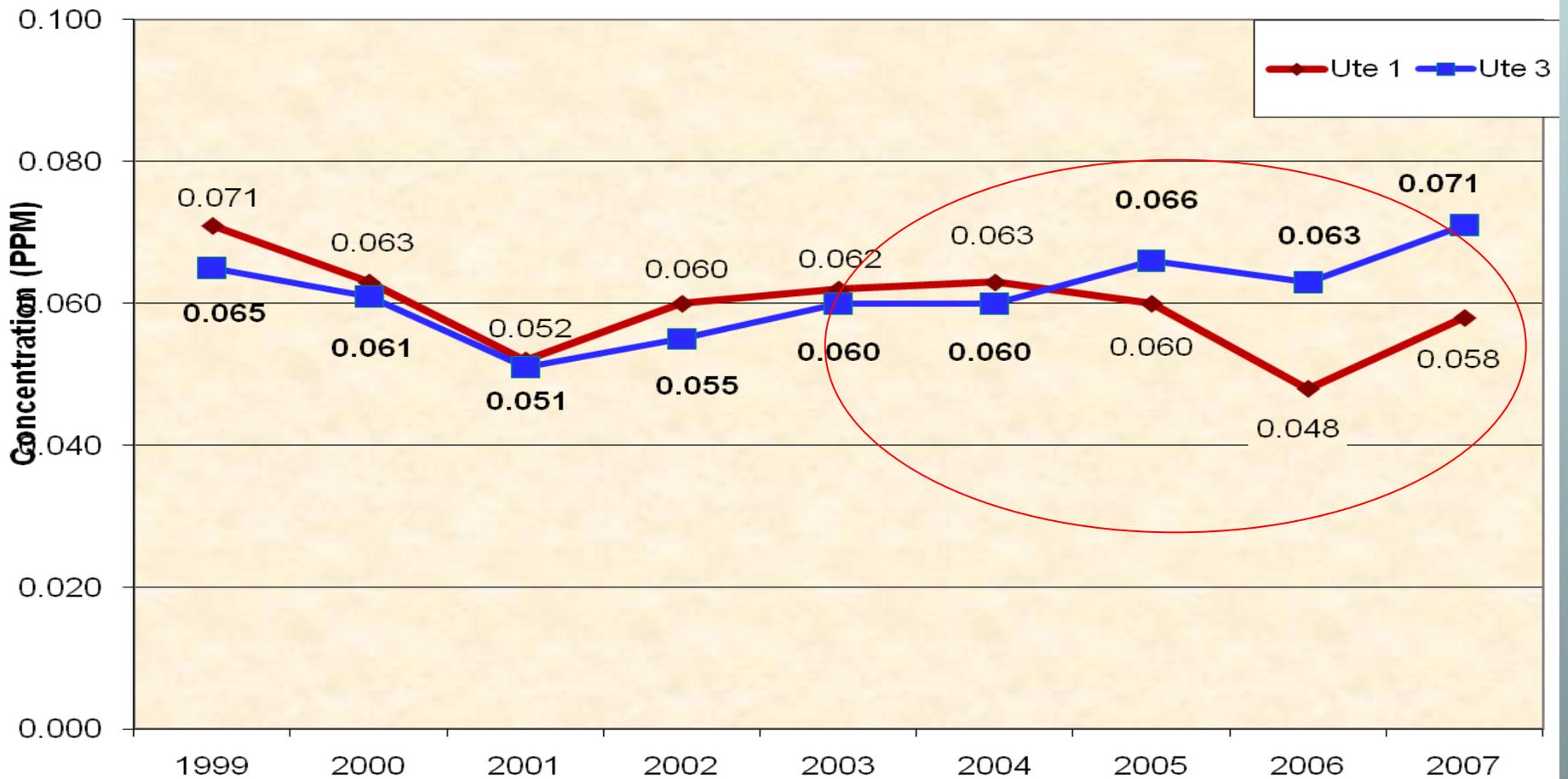
Step 5: Ask your question

- Ex: When analyzing this data, we saw a shift in how the 2 sites' O3 tracked
- During one time period, one site had markedly higher O3 levels than the other site, but the rest of the time the 2 sites agreed well

Is this significant?

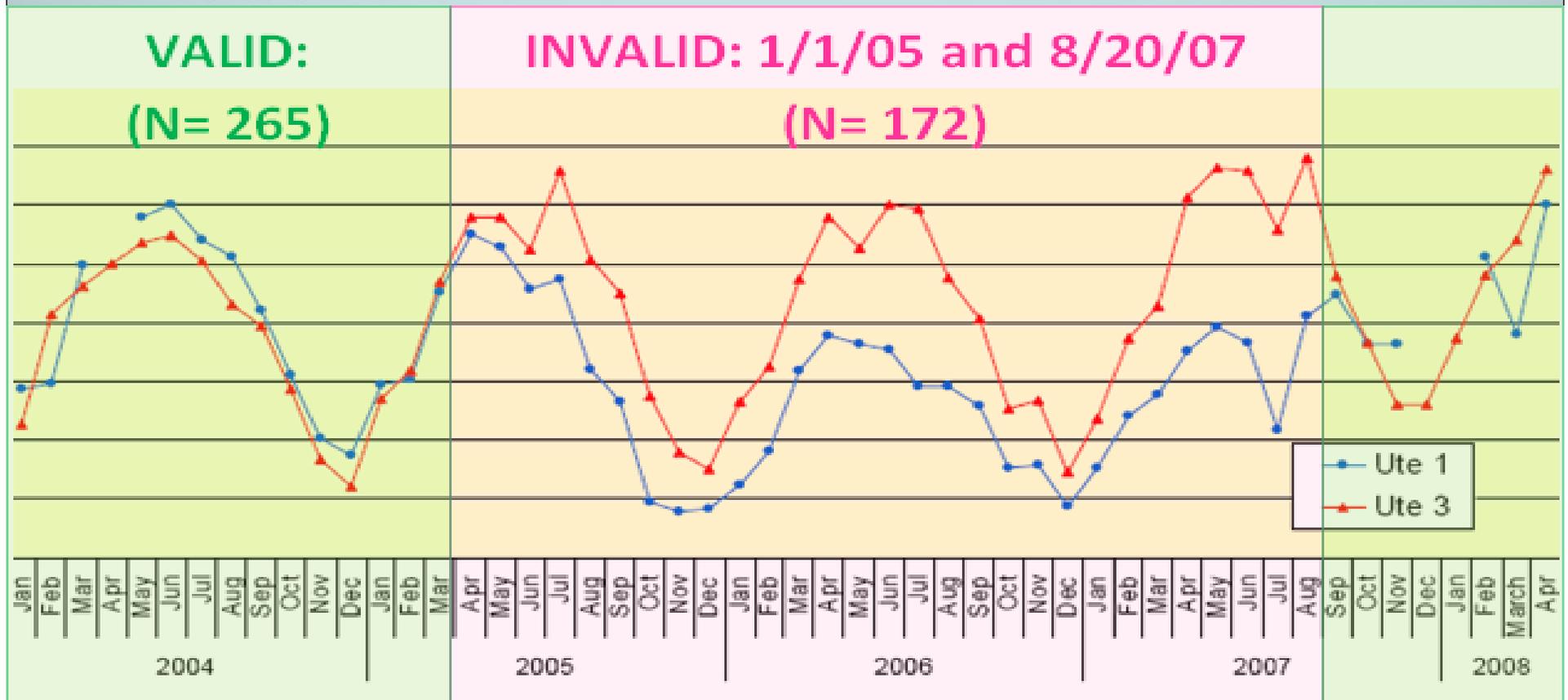
■ ---▲---Ute 1, ---■---Ute 3

8-hour Running Average Ozone 4th Max Values



Step 5: Get specific with your question:

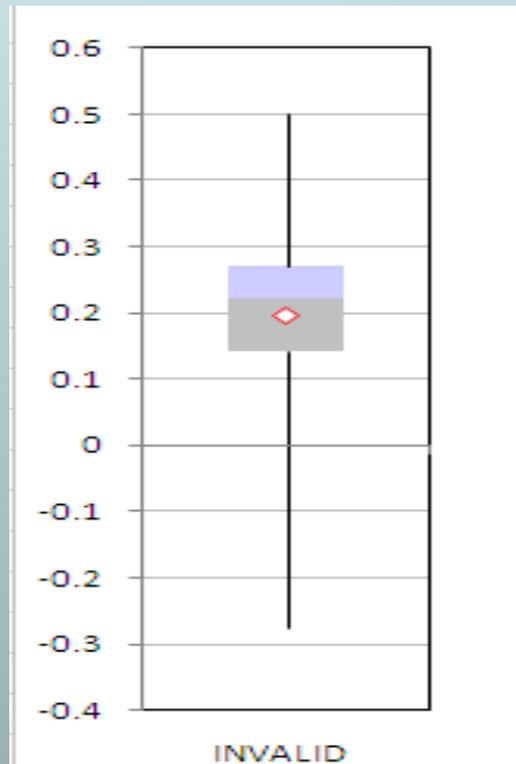
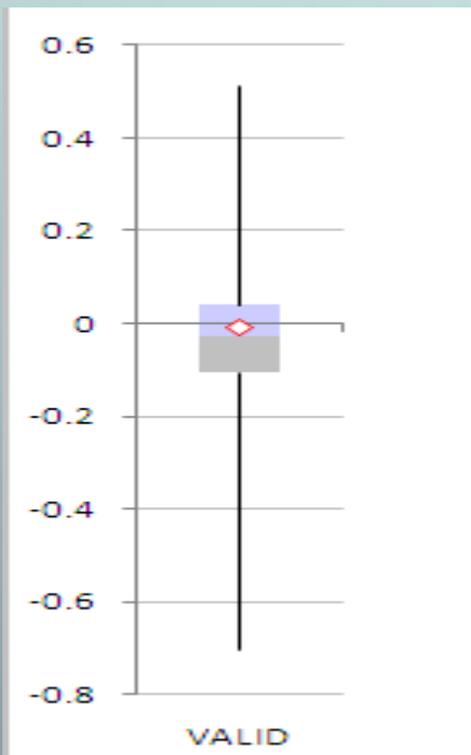
- In this case, the $(Ute\ 3 - Ute\ 1) / Ute\ 3$ ratio:



Step 6-Evaluate Distributions

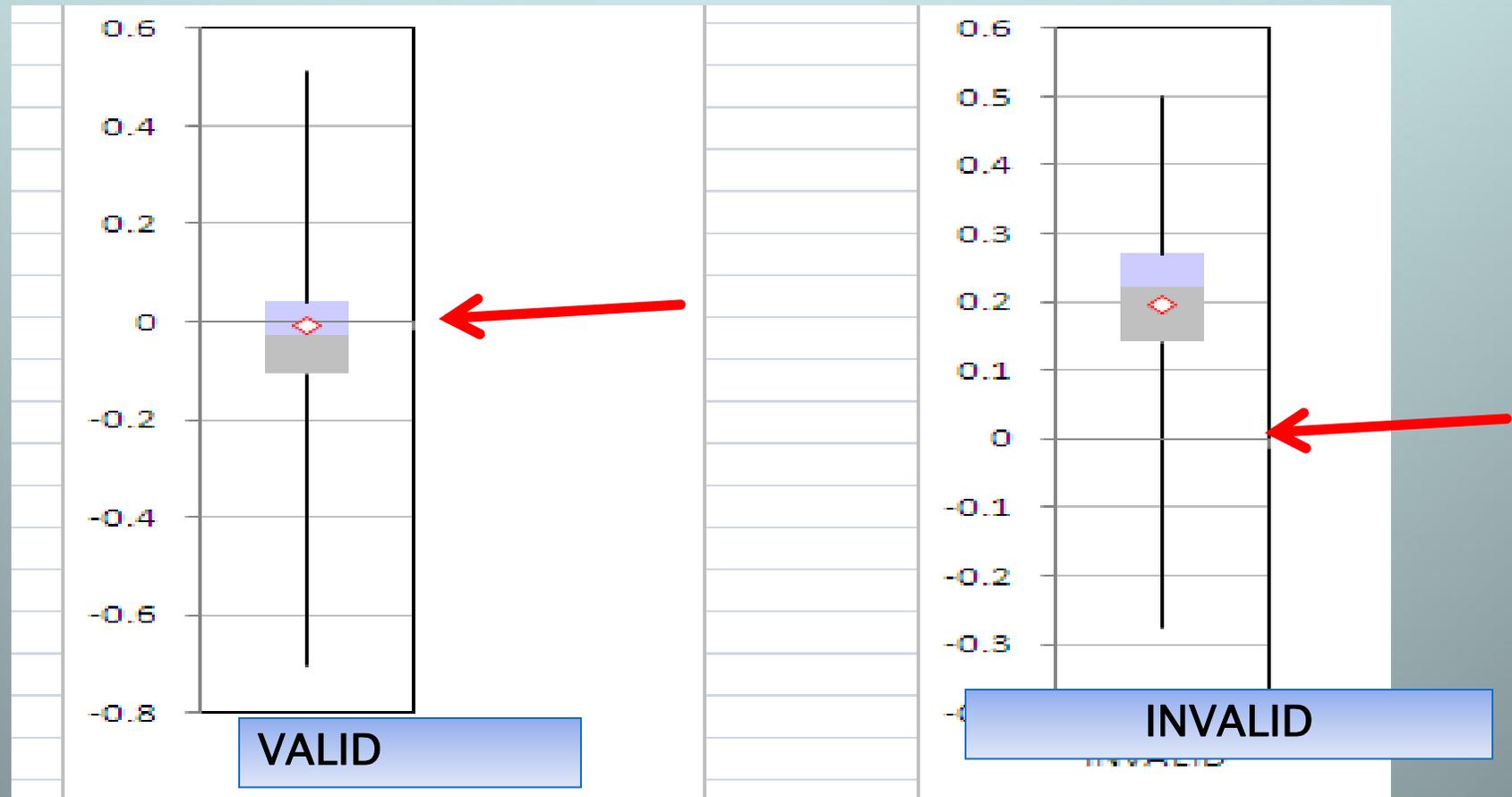
- Is this helpful?
- Sort of...

	VALID	INVALID
Mean	-0.01	0.2
Variance	0.03	0.02



Pictures more useful:

- See BoxCharter.zip for Excel Add-In to generate box and whisker plots

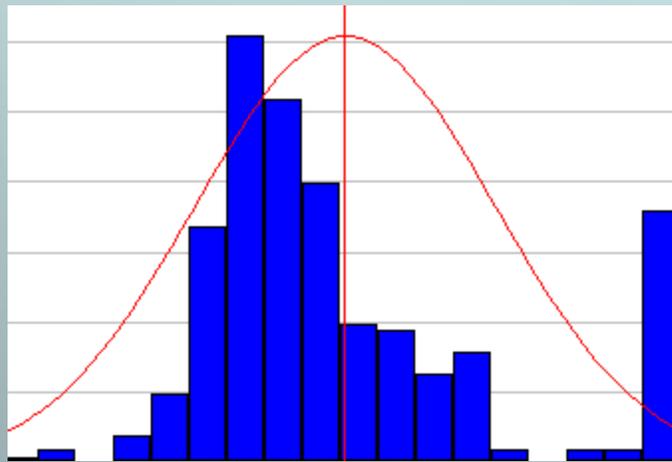


Is there a difference? How much, and how sure are we?

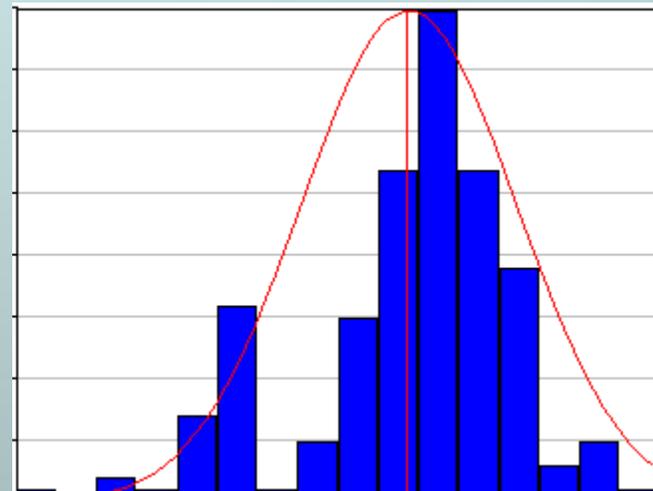
- Decide on your assumption that the data must be used to prove wrong--the null hypothesis
- If you think the data from 2 sites are different, then assume that the difference between them is zero and then the data must prove that wrong, at some level of confidence
- the null hypo is that there is zero difference between the means of the datasets

Step 6: Evaluate distributions:

- Are they normal? “approximately”?
- If so, the tests are easier



VALID

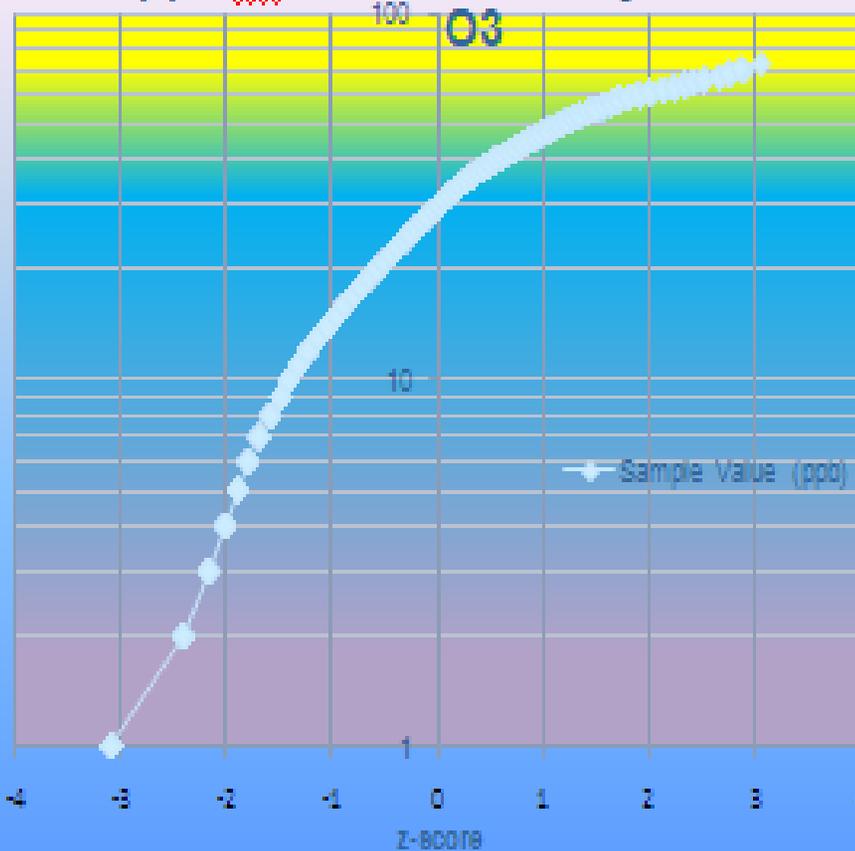


INVALID

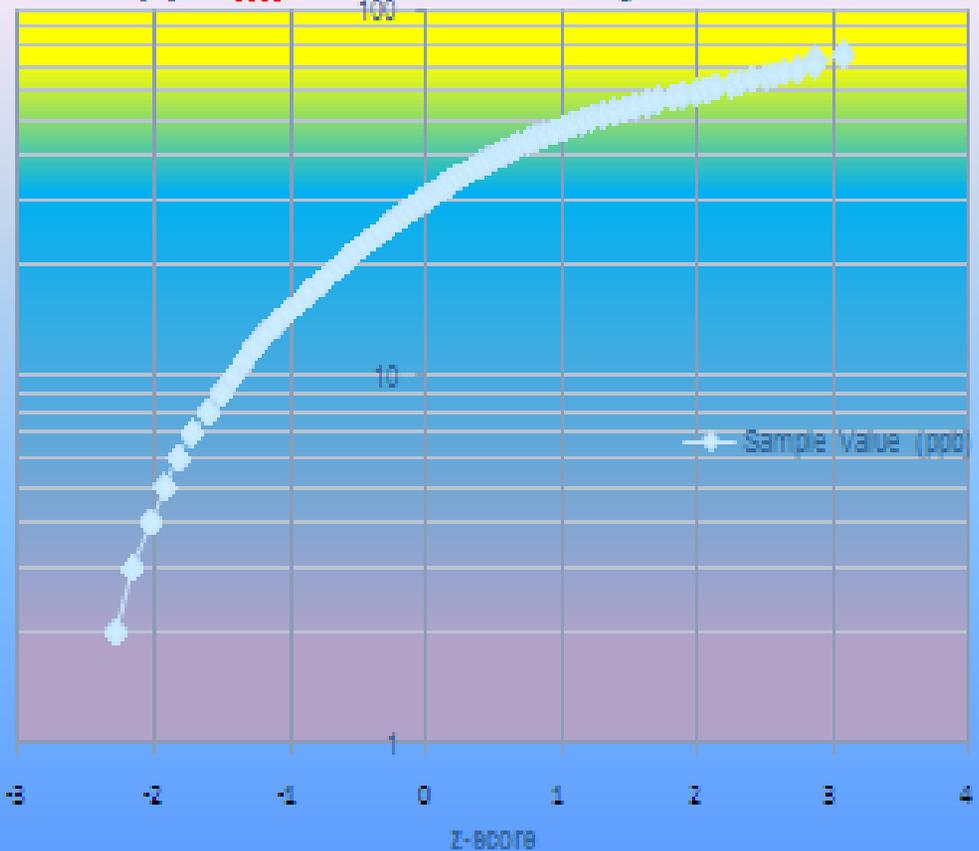
- Hmmmmm.

Use Excel Q-Q plot:

ppb vs z-score: linearity of Ute 3



ppb vs z-score: linearity of Ute 1 O3



Straight line is “perfectly normal” ...

Step 7: Use tests in Excel's Data Analysis Toolpack:

t-Test: Two-Sample Assuming Unequal Variances:

	<i>INVALID</i>	<i>VALID</i>
Mean	0.20	-0.01
Variance	0.02	0.03
Observations	172	265
Hypothesized Mean Diff	0	
t Stat	14.15	<i>NULL HYPO</i>
P(T<=t) one-tail	8.01E-38	<i>IS EXTREMELY</i>
t Critical one-tail	1.65	<i>UNLIKELY</i>

Conclusion:

- The “invalid” dataset was removed from AQS
- New audits were conducted, a 2nd analyzer was collocated with Ute 1, and now data from that site are deemed “good”
- Southern Ute Indian Tribe is very careful with all data, and this story shows how good QC and careful data analysis yields confidence in decisions

Knowledge is power—please share!
Melinda.ronca-battista@nau.edu

Address



<http://www4.nau.edu/itep/resources/>



ITEP

institute for tribal environmental professionals



RESOURCES

Data Analysis

AAA Data Analysis and Interpretation folders

Home » Resources

ITEP Resources:

The resource center is a clearinghouse and point of contact for information environmental staff. We have hard copy resources available as well as the access through this page. We love to hear from you, please email, call, write recommendations, or just to say hello!

