Data Quality Assessment

RYAN BROWN

US EPA REGION 4

REGION 4 AIR MONITORING QA TRAINING

SEPTEMBER 19, 2019

Overview

- Statistics Overview
- Control Charts
- Box and Whisker Plots
- Data Quality Reports (with examples of good and suspect data)
- AQS reports and tools

Goals

- Understand the types of data assessment that should be conducted by the monitoring organization and PQAO
- Be aware of the various tools available for use in evaluating data quality indicators and how to use them
- Identify data whose quality appears suspect

CFR Equations for Data Quality Assessments

Found in 40 CFR 58, Appendix A, Section 4

Equation 1

$$d_i = \frac{\text{meas} - \text{audit}}{\text{audit}} \times 100$$

Equation 2

$$CV = \sqrt{\frac{n \cdot \sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2}{n(n-1)}} \cdot \sqrt{\frac{n-1}{X^2_{0.1,n-1}}}$$

Equation 3

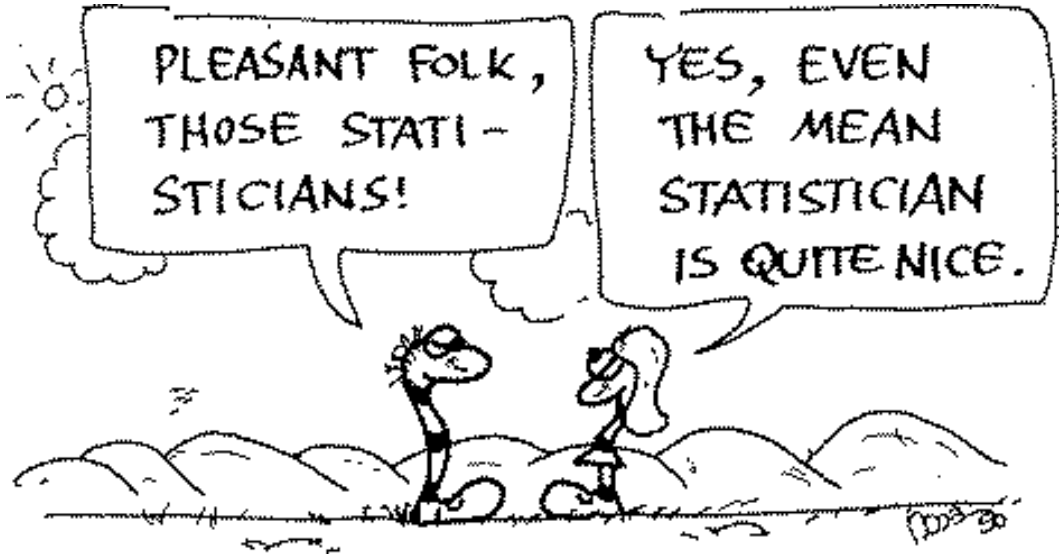
$$|AB| = AB + t_{0.95,n-1} \cdot \frac{AS}{\sqrt{n}}$$

Equation 4

$$AB = \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i|$$

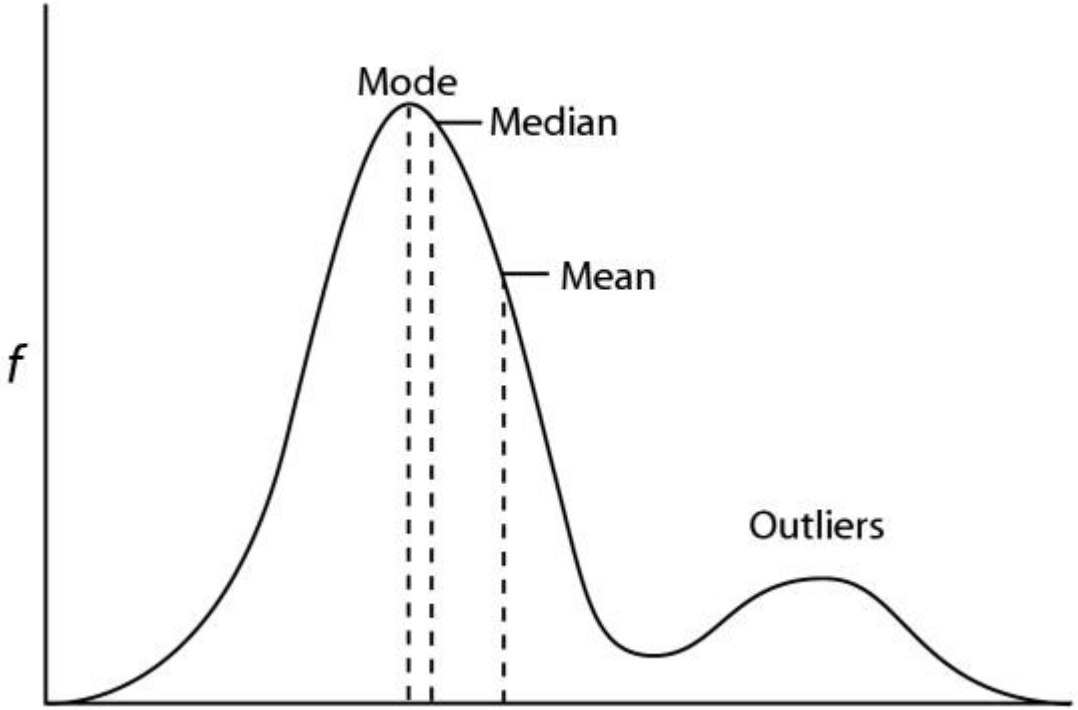
Equation 5

$$AS = \sqrt{\frac{n \cdot \sum_{i=1}^{n} |d_i|^2 - \left(\sum_{i=1}^{n} |d_i|\right)^2}{n(n-1)}}$$



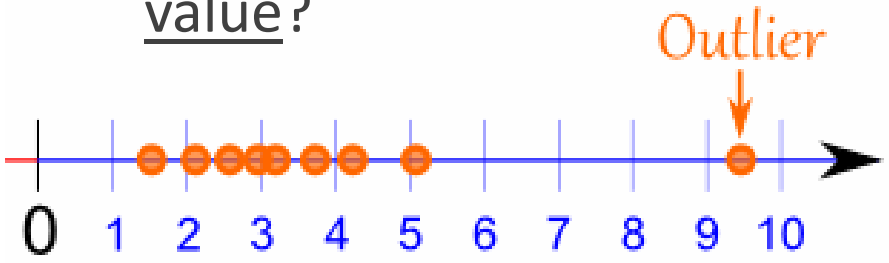
Statistical Terms

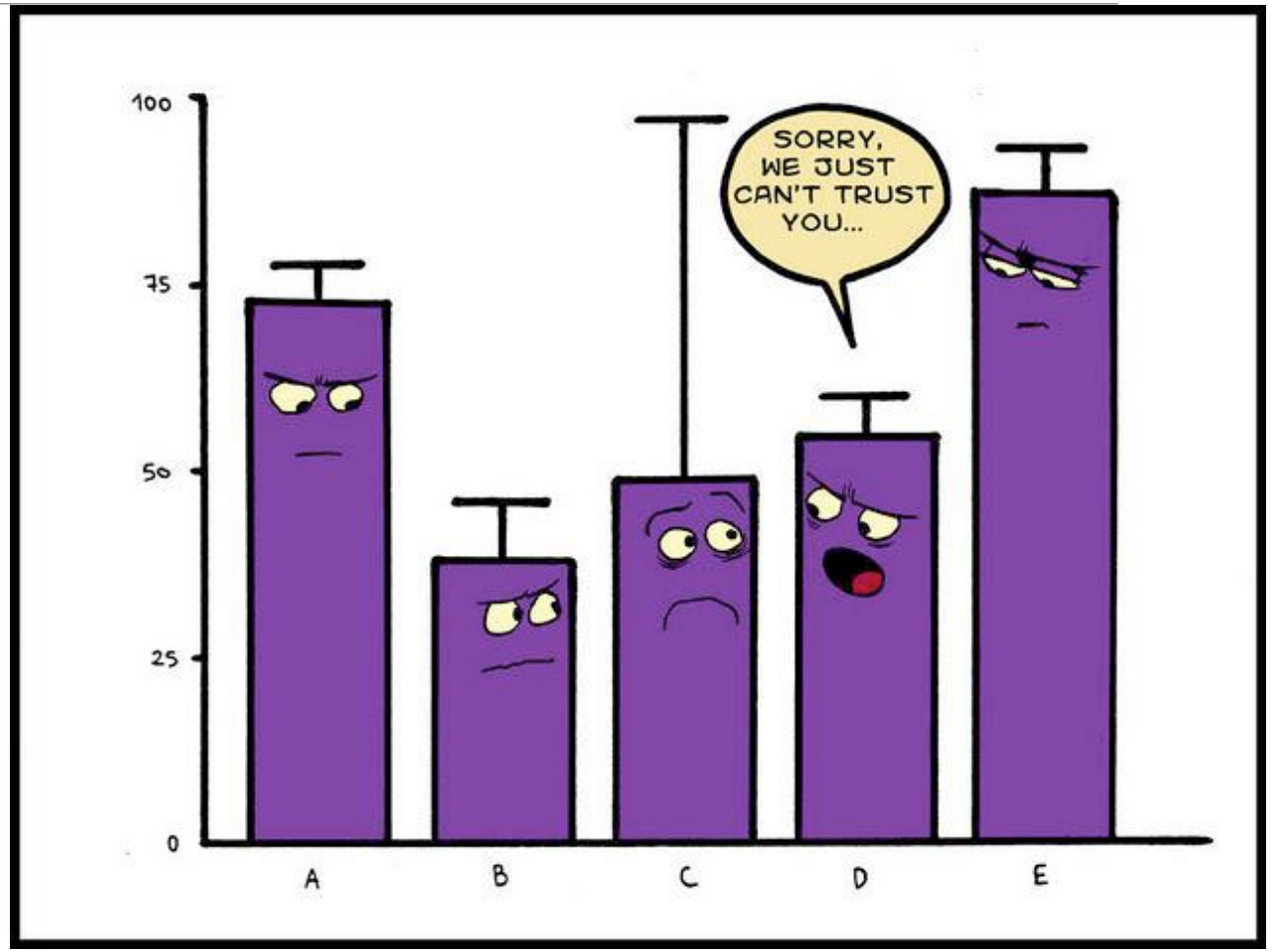
- Arithmetic Mean
- Standard Deviation
- Median
- Mode



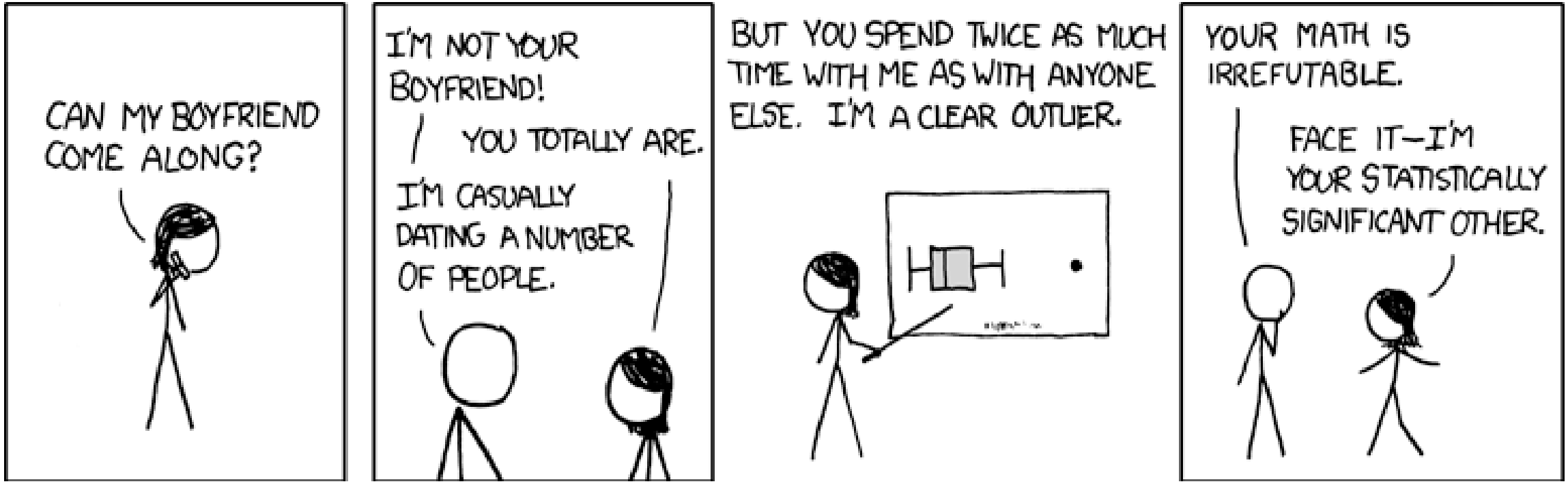
Statistical Terms

- Outlier
  - observation point that is distant from other observations
  - Is this <u>measurement error</u> or an actual <u>high/low measured value</u>?





http://andrewgelman.com/2011/12/suspicious-histograms/



https://xkcd.com/539/

Data Assessment can be useful in other ways

OR FOR AIR MONITORING…

Calculations of Data Quality Assessments: 40 CFR Part 58 Appendix A, Section 4

Provides explanation and formulas for a variety of statistical measures of uncertainty related to air quality data

AQS standard reports and other EPA tools use these formulas

You can perform these calculations on your own, or use the EPA Data Assessment Statistical Calculator (DASC)
http://www.epa.gov/ttn/amtic/qareport.html

Relative difference between measurements:

Equation 1

$$d_i = \frac{\text{meas} - \text{audit}}{\text{audit}} \times 100$$

Percent Difference
(for audit comparison)

Equation 6

$$d_i = \frac{X_i - Y_i}{(X_i + Y_i)/2} \cdot 100$$

Precision Estimate
(for collocated samplers)

EPA Statistics vs Math Statistics:

	CFR/EPA	**Statistics**
$d_i = \frac{\text{meas} - \text{audit}}{\text{audit}} \times 100$	"percent difference" a comparison of an audit concentration or value to the concentration/value measured by the monitor	"percent error" the difference between a measured and known value, divided by the known value, multiplied by 100%
$d_i = \frac{X_i - Y_i}{(X_i + Y_i)/2} \cdot 100$	"collocated precision estimate" Precision is estimated via duplicate measurements from collocated samplers	"percent or relative difference" calculated when you want to know the difference in percentage between two numbers.

Bias - 40 CFR 58 Appendix A 4.1.3

A statistic is **biased** if it is calculated/measured in such a way that it is systematically different from the population parameter of interest.

Equation 3

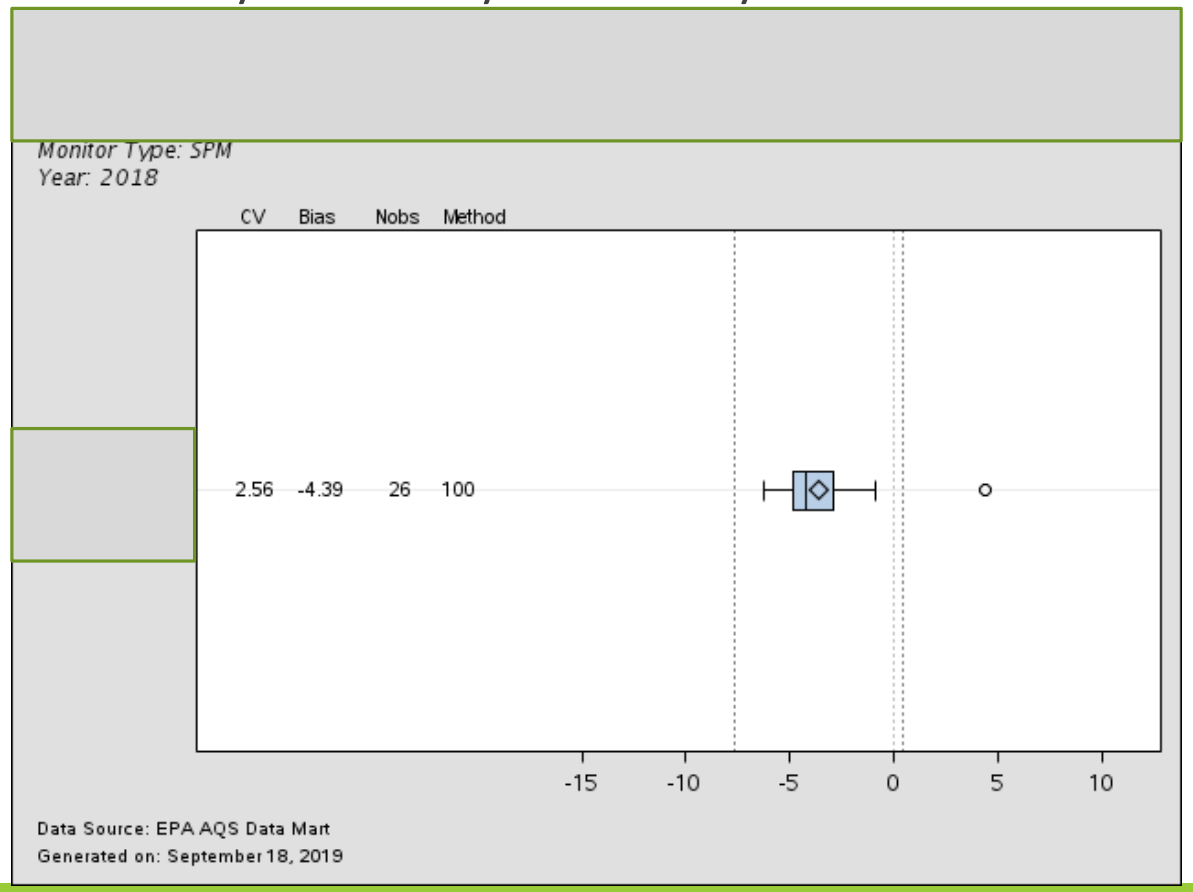
$$|bias| = AB + t_{0.95,n-1} \cdot \frac{AS}{\sqrt{n}}$$

Equation 4

$$AB = \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i|$$

Equation 5

$$AS = \sqrt{\frac{n \cdot \sum_{i=1}^{n} |d_i|^2 - \left(\sum_{i=1}^{n} |d_i|\right)^2}{n(n-1)}}$$



Coefficient of Variation

In probability theory and statistics, the **coefficient of variation** (CV) is a normalized measure of dispersion of a probability distribution or frequency distribution. It is defined as the ratio of the standard deviation to the mean.

Equation 2

$$CV = \sqrt{\frac{n \cdot \sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2}{n(n-1)}} \cdot \sqrt{\frac{n-1}{\chi^2_{0.1,n-1}}}$$

CV for audit comparison

Equation 7

$$\mathrm{CV} = \sqrt{\frac{\mathrm{n} \cdot \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{d}_\mathrm{i}^2 - \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{d}_\mathrm{i}\right)^2}{2\mathrm{n}(\mathrm{n}-1)}} \cdot \sqrt{\frac{\mathrm{n}-1}{\mathrm{X}^2_{0.1,\mathrm{n}-1}}}$$

CV for collocated samplers

Coefficient of Variation

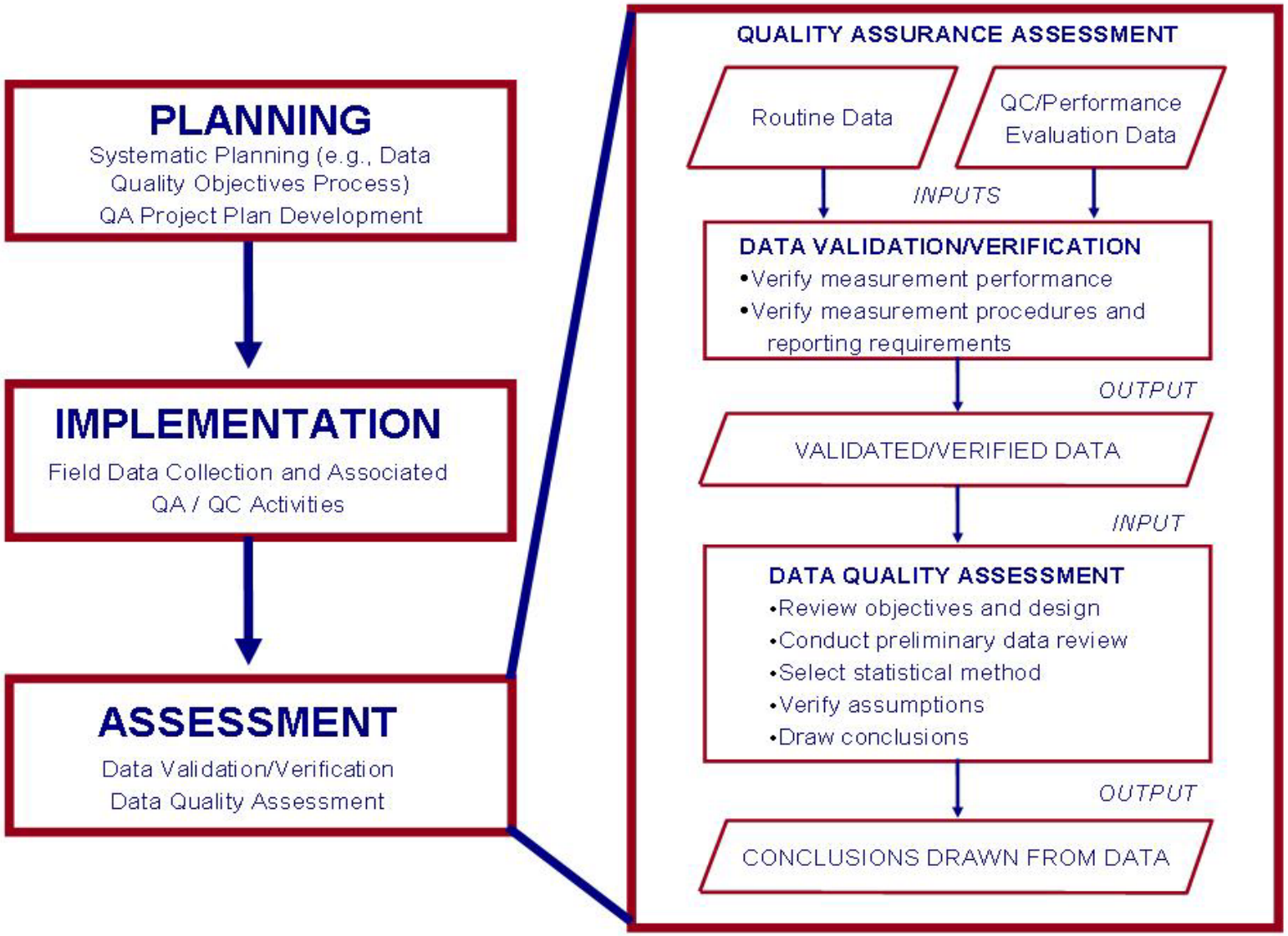
Advantage

- Where as, the standard deviation of data must always be understood in the context of the mean of the data.
- In contrast, the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number.
- For comparison between data sets with different units or widely different means (such as monitoring sites across a network), one should use the coefficient of variation instead of the standard deviation.

Data Quality Assessments and Data Certified at the PQAO level

"The monitoring organization identified as the PQAO will be responsible for the **oversight** of the quality of data of all monitoring organizations within the PQAO"

40 CFR Part 58 Appendix A Sec 1.2.1



Data Quality Assessment

The process of evaluating data against the Data Quality Objectives (DQOs)

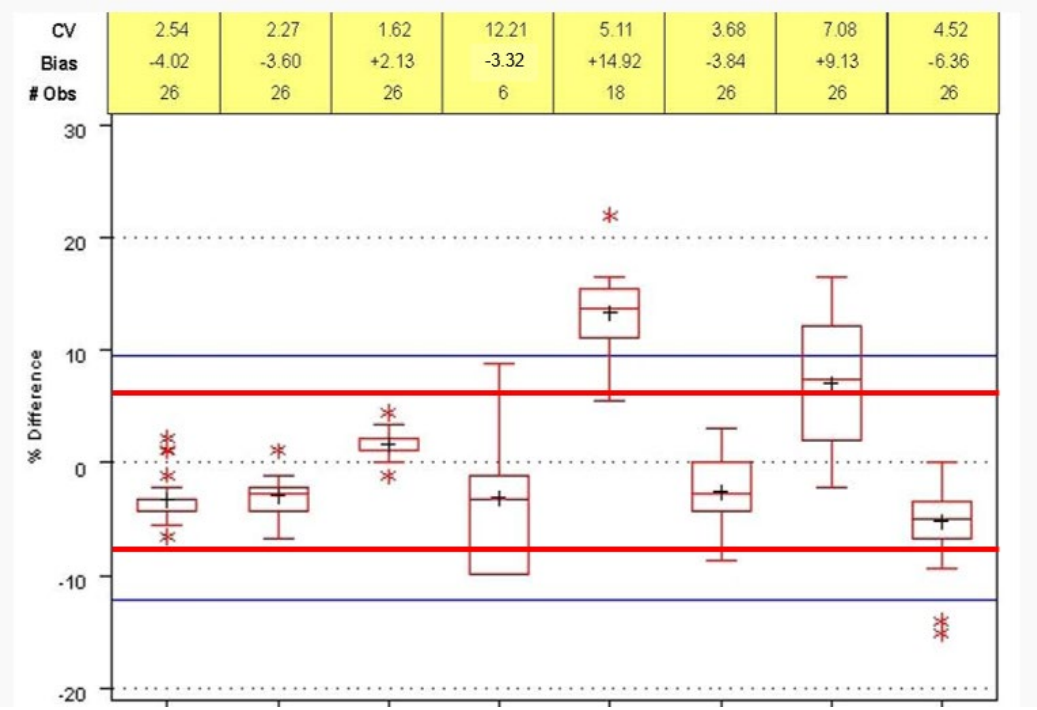
Verification Vs. Validation Vs. Assessment

- **Data Verification** - the process of evaluating the completeness, correctness, and conformance /compliance of data against the method, procedural, or contractual requirements
  - Within the monitoring org.
  - Usually by field/lab personnel
  - Smaller data sets

- **Data Validation** - to confirm, through provision of objective evidence, that particular requirements for a specified intended use are fulfilled. Extends the evaluation of data beyond method, procedural, or contractual compliance.
  - Within the monitoring org.
  - Usually by field/lab manager or QA manager
  - Larger data sets
  - Performed prior to reporting to AQS (quarterly)

- **Data Quality Assessment** - the process to determine if the data are suitable for a specific use over the DQO data aggregation time period
  - Performed within and outside monitoring org
  - Can be performed on larger data sets related to DQO decisions

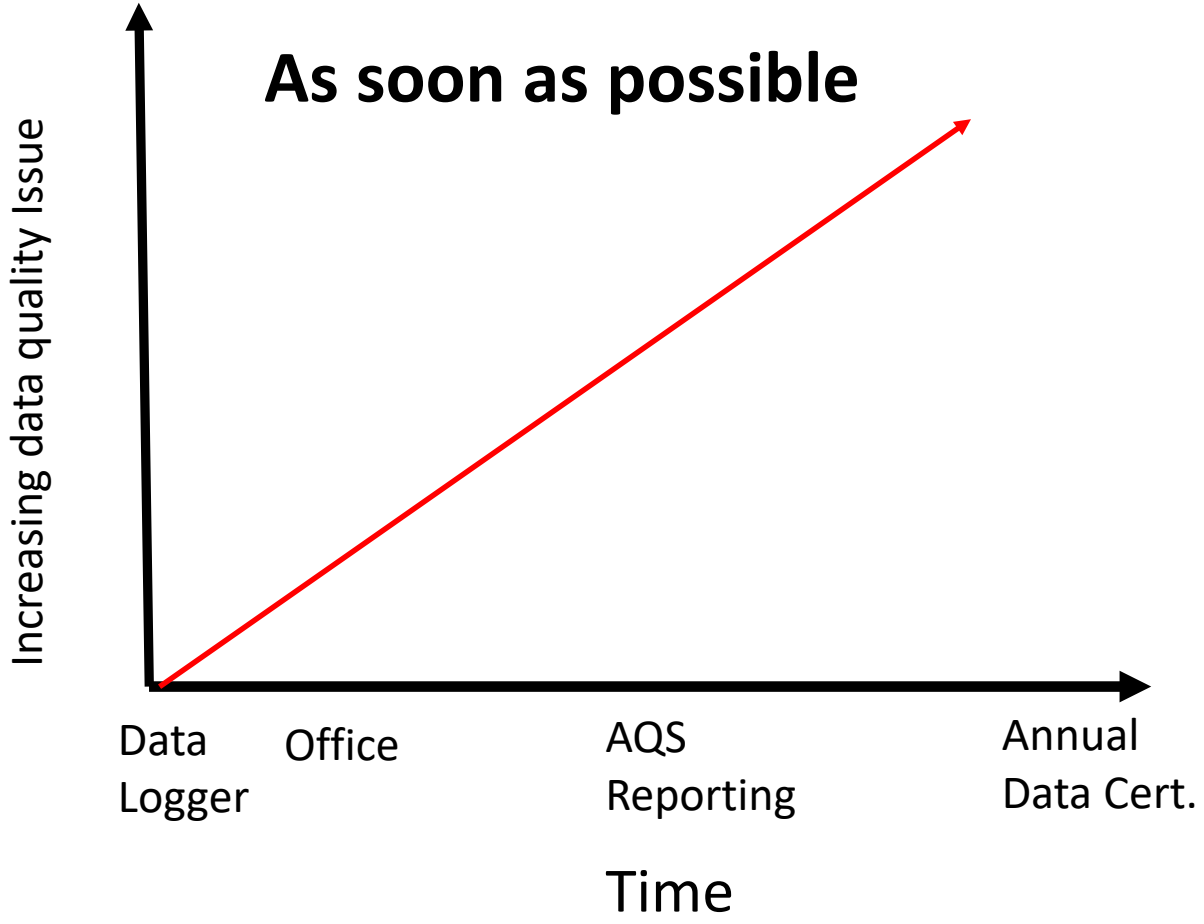
Data Assessments



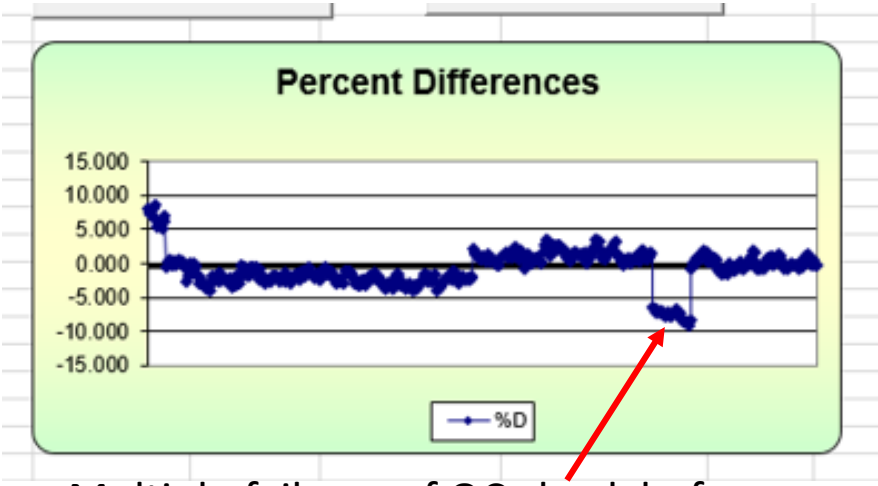
Annual Box & Whisker Plots – PQAO Level

- Monitor-level and network-level (PQAO)
- Annual data assessments should be completed by QAM (or other designated staff)
- Compare data to DQOs
- Part of data certification
- 3-year assessments are also helpful when assessing criteria pollutant data
- Longer-term assessments (e.g., 6-year or 10-year) may happen in some programs, like toxics

When the Best Time to Perform Assessments?

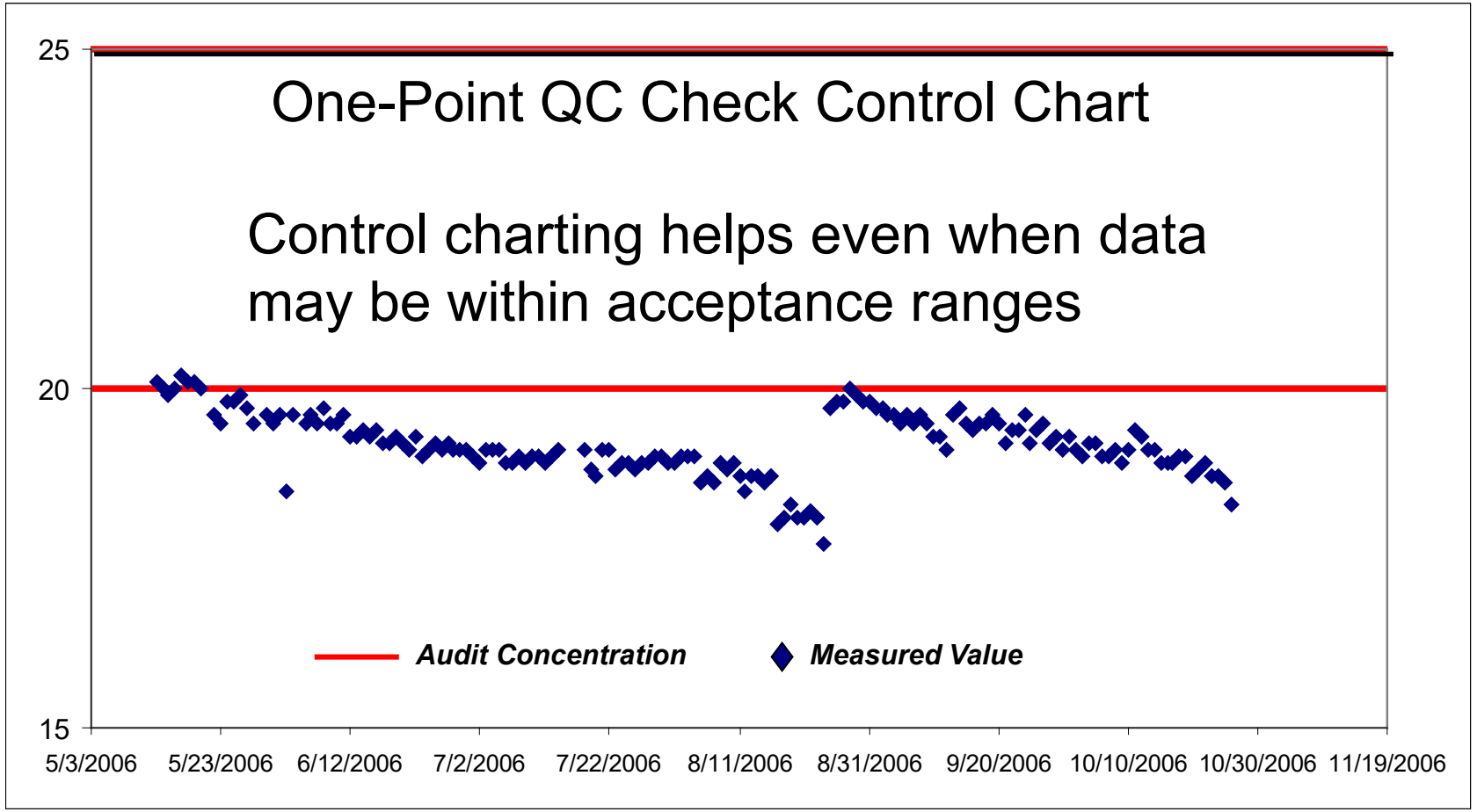


- Don't wait till data is in AQS
- Exceedances can multiply causing more data to be invalidated.
- Keep up with the small stuff to avoid the big problems

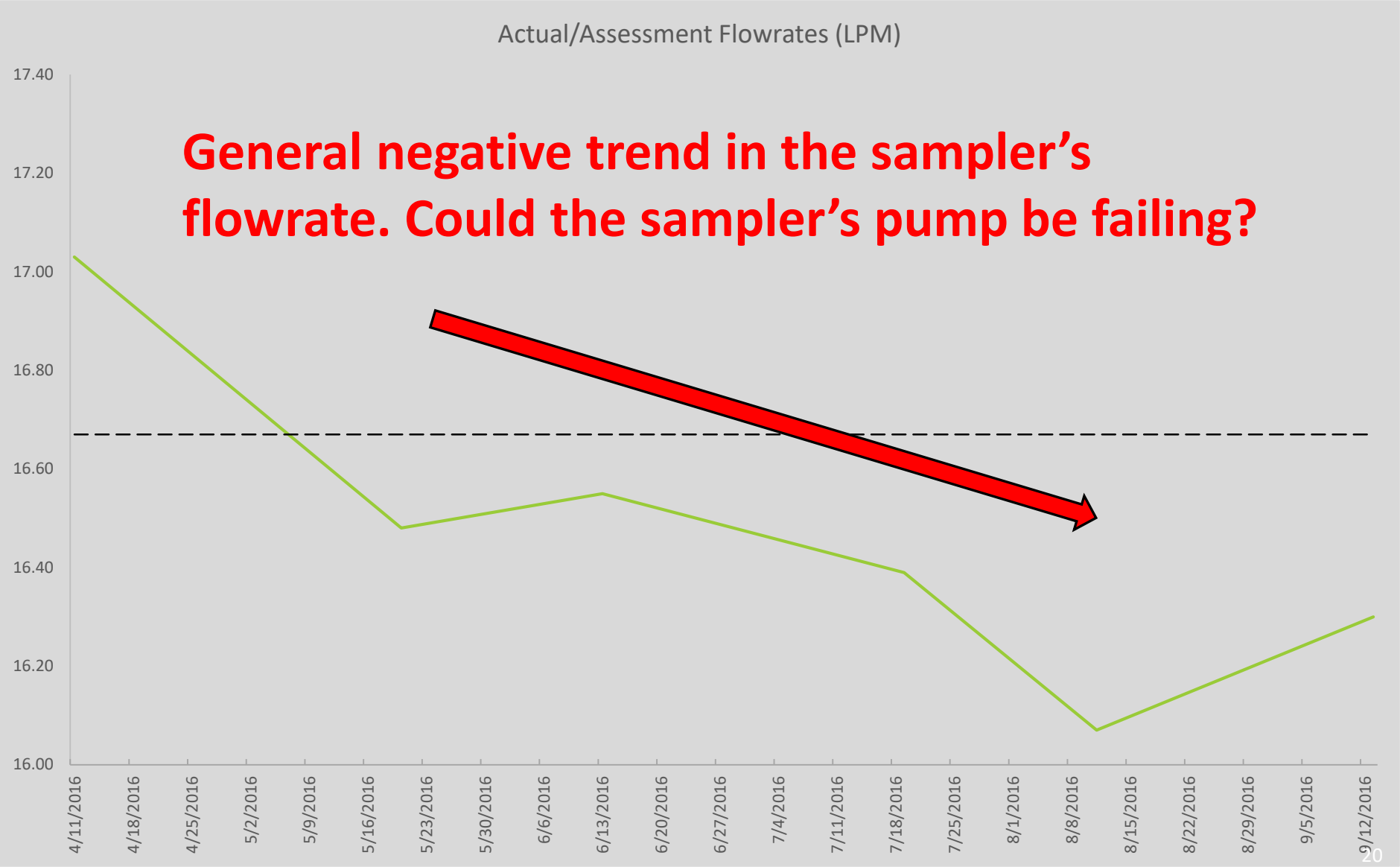


Multiple failures of QC check before corrective action…..Not good

Looking at the data



Looking at the data- PM2.5 Flowrate Control Chart

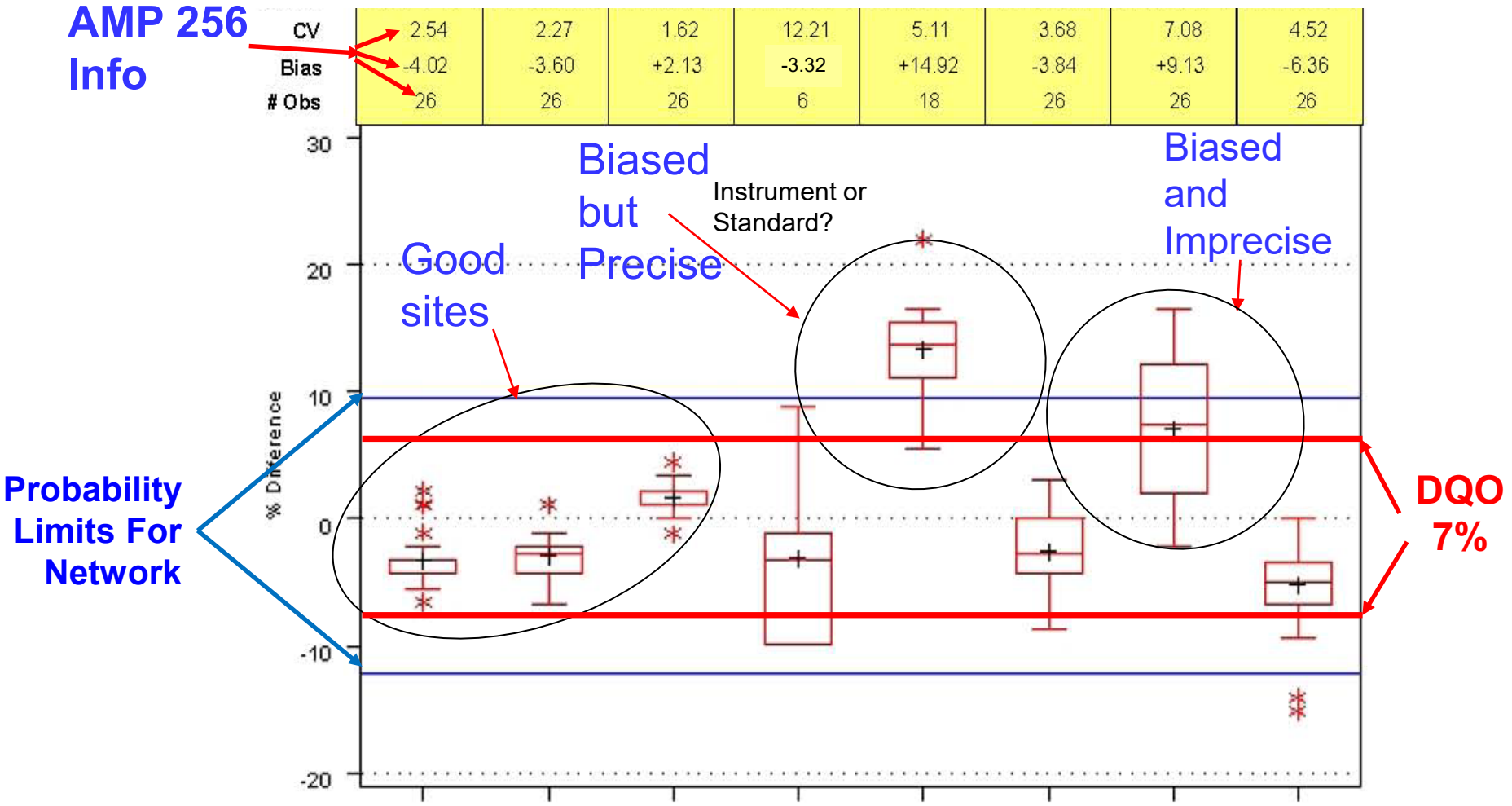


2016 flow verification check reported assessment flowrates for one sampler

Dashed line at 16.67 LPM design value

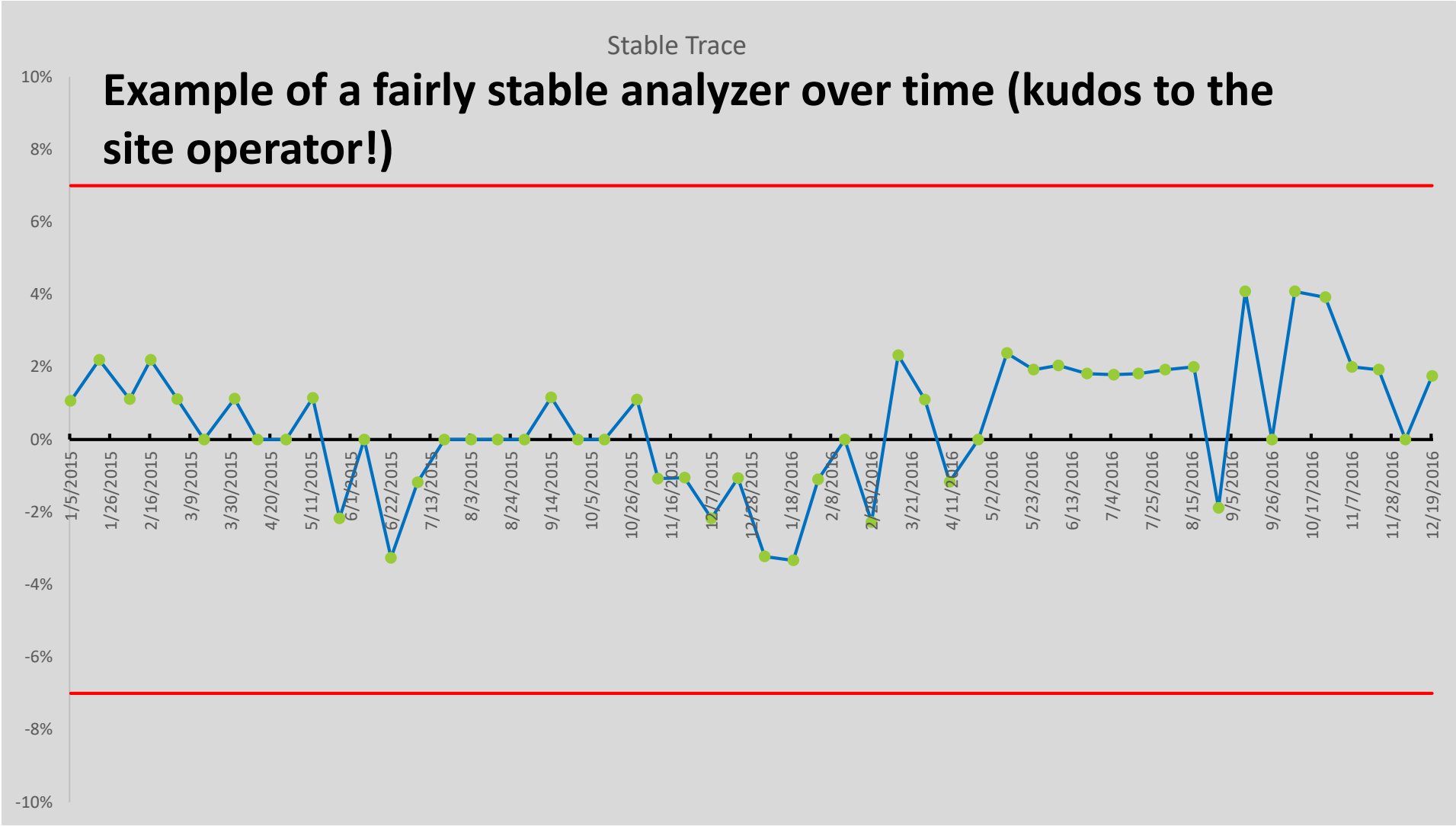
Looking at the Data

Annual Box and Whisker Plots 1 Year of Ozone Data from a Primary Quality Assurance Organization

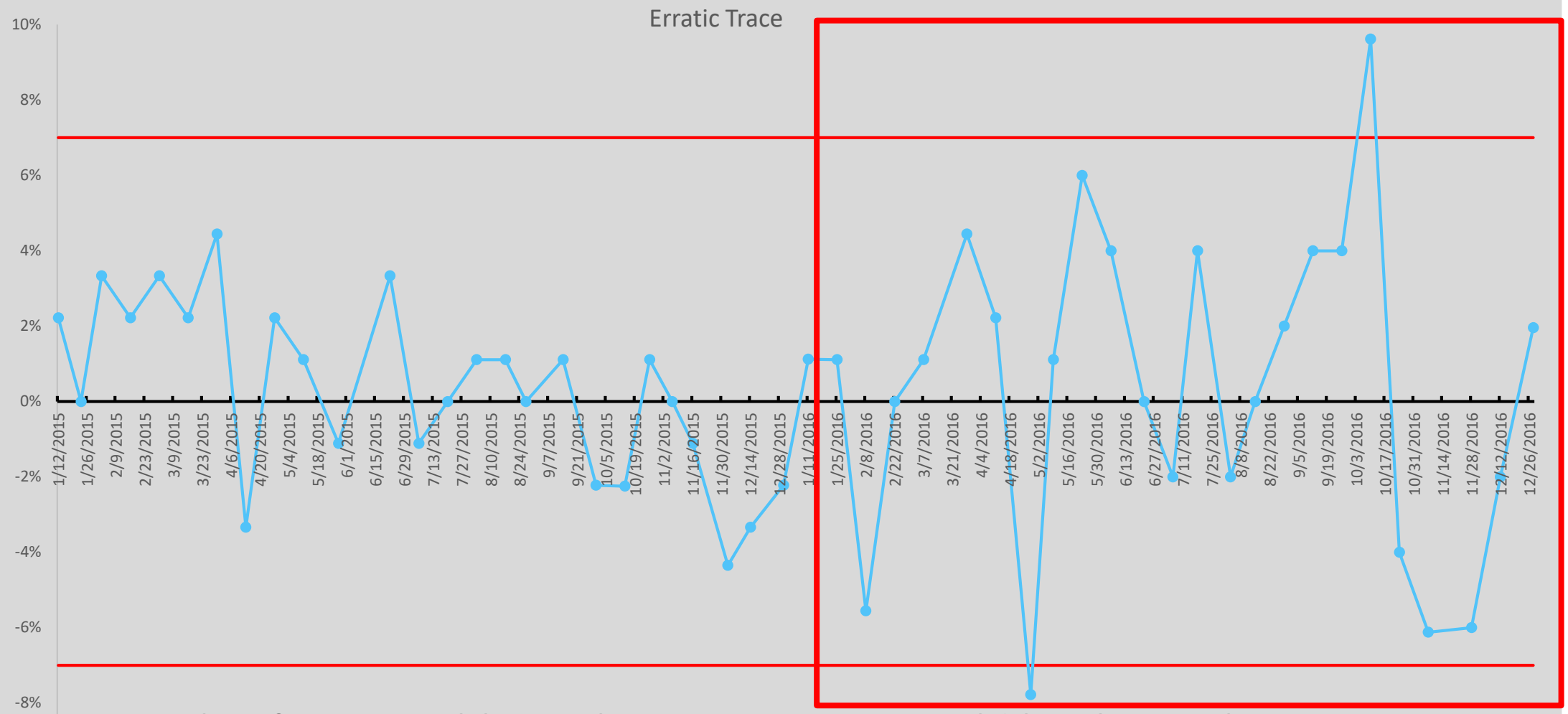


- Lets you look at all sites
- Can help identify sites in need of corrective action
- Automated application on Air Data Website

Example Control Charts (O3)



Example Control Charts (O3)



Example of an unstable analyzer over time. Could be the analyzer, site operator, or a combination

Looking at the data

- Develop in house tools to visualize your data
- Observe trends in data, develop action plans before data is out of control
- Analyze data over different timescales
- Calculate different data statistics
- Track progress toward achieving DQOs
- Field/Lab Technician: Verify QC data
- Data Validator: Ensure data are suitable for their intended use
- Monitoring Manager/Statistician: DQOs, AQS data, annual data certification

AQS Data Quality Reports

THE AMP REPORTS

AQS REPORTS

Report Category	Reports
Site and Monitor Metadata	• AMP220 (Monitor Networks) • AMP380 (Site Description) • AMP390 (Monitor Description)
Detailed Data and Summary	• AMP350 (Raw Data) • AMP350MX (Raw Data Max Values) • AMP350P • AMP350NW (Raw Data) • AMP500 (Extract Site/Monitor Data) • AMP501 (Extract Raw Data) • AMP503 (Extract Blanks Data)
Certification and QA	• AMP251 (QA Raw Assessment) • AMP256 (QA Data Quality Indicator) • AMP430 (Data Completeness) • AMP504 (QA Data Extraction) • AMP600 (Certification Evaluation and Concurrence)
Other	• AMP360 (Raw Data Qualifiers) • AMP480 (Design Value)

AMP600 Certification Evaluation and Concurrence Report

Data Evaluation and Concurrence Report Summary

Jan. 11, 2018

Certification Year: 2017
Certifying Agency [redacted]

Pollutants in Report:

Parameter Name	Code	Monitors Evaluated	Monitors Recommended for Concurrence by AQS	Monitors NOT Recommended for Concurrence by AQS
Carbon monoxide	42101	2	2	0
Lead (TSP) LC	14129	6	3	3
Nitrogen dioxide (NO2)	42602	5	4	1
Ozone	44201	34	33	1
PM10 Total 0-10um STP	81102	5	3	2

PQAOs in Report:

PQAO Name	PQAO Code	TSA Date
[redacted]		10/11/16
		04/18/16

Summary of 'N' flags for all pollutants:

PQAO	Parameter Code	AQS Site-ID	POC	AQS Recommended Flag	Cert. Agency Recommended Flag	Reason for AQS Recommendation
[redacted]	14129	[redacted]	2	N		PQAO-Level Collocation criteria not met.
	14129		1	N		PQAO-Level Collocation criteria not met.
	14129		9	N		PQAO-Level Collocation criteria not met.
	42602		1	N		Annual Performance Evaluation Audit Missing or 1 Level.
	44201		1	N		1-Point QC Precision > 20%.
	81102		4	N		Annual Summary completeness < 70%.
	81102		4	N		Annual Summary completeness < 70%.

Signature of Monitoring Organization Representative: _____

- Used in annual data certification process
- Automated flagging by AQS for any data out of acceptance criteria
- Provides a summary as well as individual pollutant evaluations
- Uses same assessment statistics used in other AMP reports

AMP600 Individual Pollutant Report(s)

PQAO Name

QAPP Approval Date 03/28/2016

NPAP Audit Summary:

Number of Passed Audits	NPAP Bias	Criteria Met
8	2.19063	Y

AQS Site ID	POC	Monitor Type	Routine Data Mean	Min	Max	Exceed. Count	Outlier Count	Perc. Comp.	One Point Quality Check Precision	Bias	Complete	Annual PE Bias	Complete	NPAP Bias	PQAO Level Criteria	QAPP Appr.	Concur. Flag Aqs Rec Flag	CA Rec Flag	Epa Concur
	1	SLAMS	0.048	0.021	0.079	0	0	100	2.33	+2.43	100	3.83	100		Y	Y	Y		
	1	SLAMS	0.047	0.015	0.075	0	0	97	1.97	+/-1.54	100	0.23	100	1.22	Y	Y	Y		
	2	SLAMS	0.042	0.002	0.072	0	0	96	2.56	-2.62	100	- 3.72	100	1.76	Y	Y	Y		
	1	SLAMS	0.049	0.018	0.087	0	0	99	2.07	+/-1.71	100	- 1.44	100		Y	Y	Y		
	1	SLAMS	0.048	0.013	0.091	0	0	81	2.40	+/-2.06	100	- 0.29	100	7.67	Y	Y	Y		
	1	SLAMS	0.052	0.020	0.094	0	0	98	2.61	+2.35	100	0.31	100		Y	Y	Y		
	1	SLAMS	0.050	0.017	0.086	0	0	90	1.32	+2.16	100	- 1.90	100		Y	Y	Y		
	2	SLAMS	0.050	0.019	0.084	0	0	99	2.11	+4.48	100	4.09	100	2.81	Y	Y	Y		
	1	SLAMS	0.049	0.018	0.076	0	0	100	3.74	+4.08	100	- 10.04	100	3.91	Y	Y	Y		
	1	SLAMS	0.040	0.022	0.069	0	0	100	31.35	+/- 11.35	100	0.12	100		Y	Y	N		
	1	SLAMS	0.051	0.028	0.082	0	0	99	1.63	-1.87	100	- 1.59	100		Y	Y	Y		

AQS DOCUMENTATION

https://www.epa.gov/aqs

https://www.epa.gov/aqs/aqs-manuals-and-guides

https://aqs.epa.gov/aqsweb/documents/codetables/certification_flags.html

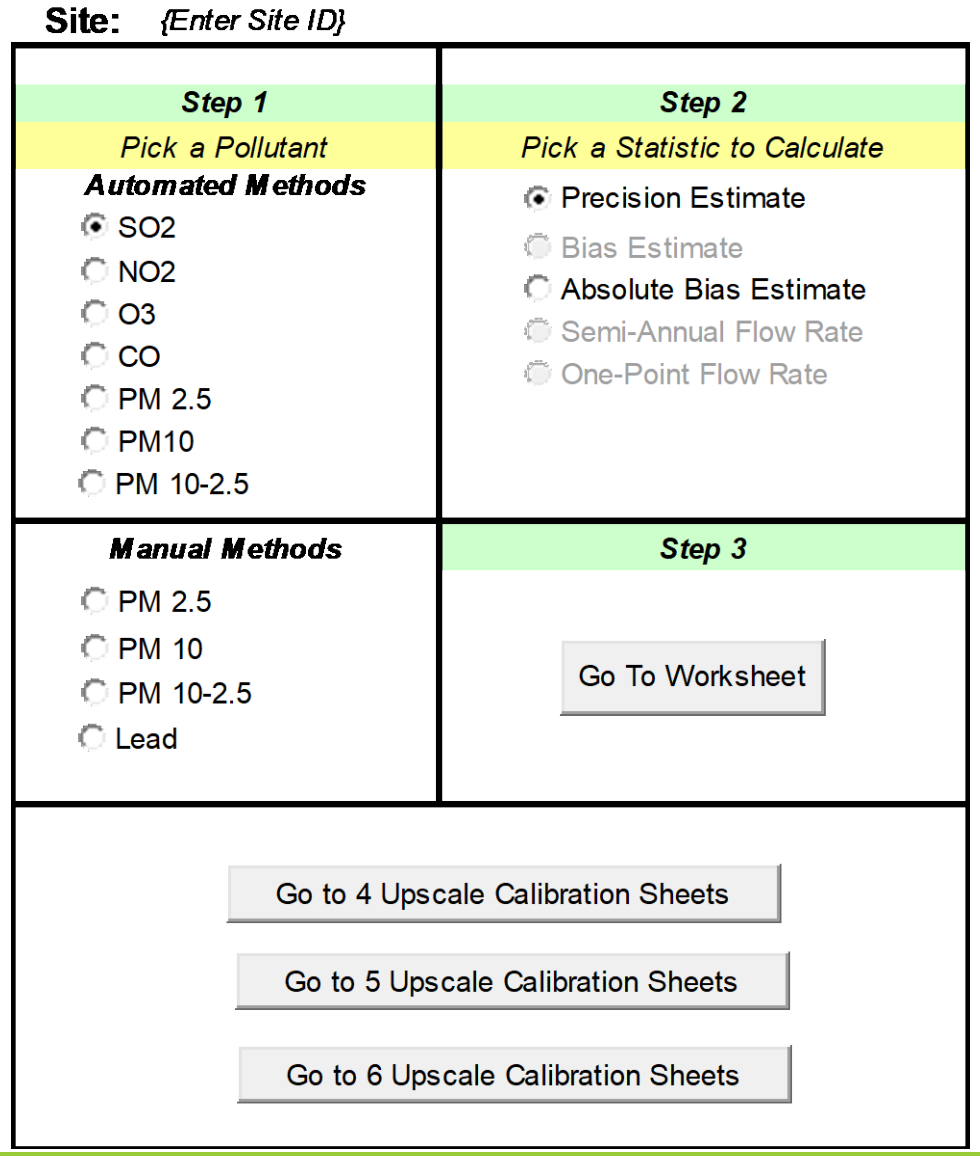
https://www.epa.gov/sites/production/files/2017-03/documents/aqs_data_dictionary.pdf

Data Quality Assessment Tools

Tools for Data Assessment

- **Data Assessment Statistical Calculator (DASC)**
  - AMTIC https://www3.epa.gov/ttn/amtic/qareport.html
- **AQS 504 Report Assessment Tool**
  - AMTIC https://www3.epa.gov/ttn/amtic/qareport.html
- **Single Point Precision and Bias Report – (Box and Whisker)**
  - Air Data https://www.epa.gov/outdoor-air-quality-data/single-point-precision-and-bias-report
- **R-Shiny Automated PM2.5 FRM Data Quality Assessment**
  - STI Web https://sti-r-shiny.shinyapps.io/QVA_Dashboard/
- **AQS API**
  - RESTful query service https://aqs.epa.gov/aqsweb/documents/data_api.html
- **PM2.5 Comparability**
  - Continuous vs FRM comparability https://www.epa.gov/outdoor-air-quality-data/pm25-continuous-monitor-comparability-assessments

Data Assessment Statistical Calculator (DASC)



- Excel file on AMTIC!
- Menu driven!
- Provides same statistics as AMP256 Report!
- Don't have to wait till data is in AQS!
- Provide graphics of results!
- Include calibration feature!

Website

https://www3.epa.gov/ttn/amtic/qareport.html

DASC For Ozone 1-point QC Check

- Just enter audit value (X) and measured value (Y)
- The statistics are calculated on the fly
- Values will equal the AMP256 Report as long as the data is the same
- Provides graphics to help identify exceedances
- Provide CFR CV and bias statistic

O_3 Assessments

Site ID: {Enter Site ID} | Pollutant type: O_3

Meas Val (Y)	Audit Val (X)	d (Eqn. 1)	25th Percentile	d^2	|d|	$\|d\|^2$
75.7	70	8.143	-2.300	66.306	8.143	66.306
75	70	7.143	75th Percentile	51.020	7.143	51.020
75.2	70	7.429	0.677	55.184	7.429	55.184
75	70	7.143		51.020	7.143	51.020
75.7	70	8.143		66.306	8.143	66.306
76	70	8.571		73.469	8.571	73.469
74.5	70	6.429		41.327	6.429	41.327
73.7	70	5.286		27.939	5.286	27.939
74.2	70	6.000		36.000	6.000	36.000
73.9	70	5.571		31.041	5.571	31.041
73.9	70	5.571		31.041	5.571	31.041
73.5	70	5.000		25.000	5.000	25.000
74.8	70	6.857		47.020	6.857	47.020
74.4	70	6.286		39.510	6.286	39.510
73.5	73.9	-0.541		0.293	0.541	0.293
74	73.9	0.135		0.018	0.135	0.018
74.3	73.9	0.541		0.293	0.541	0.293
73.8	73.9	-0.135		0.018	0.135	0.018
73.7	73.9	-0.271		0.073	0.271	0.073
74.1	73.9	0.271		0.073	0.271	0.073
74.1	73.9	0.271		0.073	0.271	0.073
73.7	73.9	-0.271		0.073	0.271	0.073
74	73.9	0.135		0.018	0.135	0.018
74.1	73.9	0.271		0.073	0.271	0.073
74.2	73.9	0.406		0.165	0.406	0.165
74.1	73.9	0.271		0.073	0.271	0.073
74	73.9	0.135		0.018	0.135	0.018
73.5	73.9	-0.541		0.293	0.541	0.293
73.4	73.9	-0.677		0.458	0.677	0.458
72	73.9	-2.571		6.610	2.571	6.610
72.5	73.9	-1.894		3.589	1.894	3.589
72.6	73.9	-1.759		3.095	1.759	3.095
73.9	73.9	0.000		0.000	0.000	0.000
73.3	73.9	-0.812		0.659	0.812	0.659
73.8	73.9	-0.135		0.018	0.135	0.018

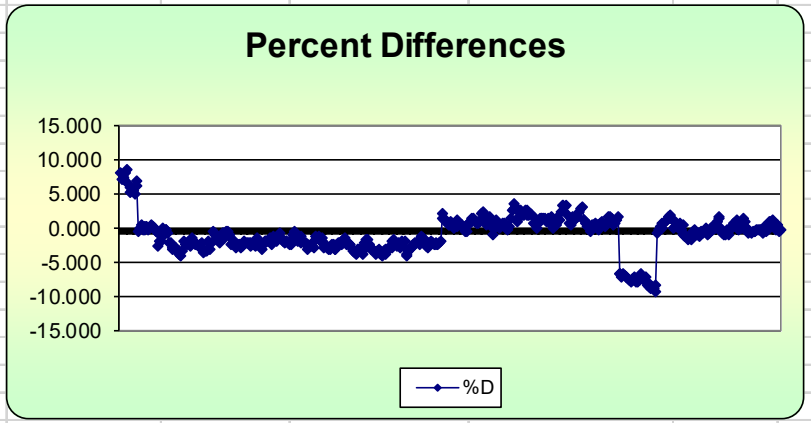
CV_{ub} (%)	Bias (%)

n	S_d	S_{d2}	$\Sigma\|d\|$	"AB" (Eqn 4)
500	2.633	15.148	1012.937	2.026
n-1	**Σd**	**Σd^2**	**$\Sigma\|d\|^2$**	**"AS" (Eqn 5)**
499	-446.201	3.858.540	3858.540	1.903

CV (%) (Eqn 2)	Bias (%) (Eqn 3)	Both Signs Positive
2.75	2.17	FALSE
	Signed Bias (%)	Both Signs Negative
	+/-2.17	FALSE

Upper Probability Limit	Lower Probability Limit
4.27	-6.05

Return to Main Menu | Print Worksheet



DASC Calibration Function

Multipoint Verification/Calibration Assessment

What percentage is acceptable?	2%
Calibration Scale	200
Point Difference Acceptance Value	1.5
Only values on sheet that can be changed are in colors	Gray Green

Instructions

1) Place calibration values (X) in gray cells in row 17
2) Enter the conc. values (Y) from the instrument in the green fields. The worksheet allows for 7 values per conc., but you can place only one row of data into the worksheet and delete the data in the other rows.
3) the remainder of the worksheet should automatically calculate the results.
4) Any point result > the point difference acceptance criteria in B8 will turn the boxes and font in rows 27 red. Any percent difference > then value in B6 will turn the boxes and font in rows 28 red.
5) The percent difference estimates are measured using the best fit conc. values and the average of the instrument values for each conc.

	Zero Concentration	Concentration 2	Concentration 3	Concentration 4	Concentration 5
Calibrator Value (X)	0	50	100	150	200
Instrument Value (Y)	0	47	99	148	203
	-1	48	98	151	200
	0	49	101	152	195
	0	46	99	152	199
	0.5	52	101	154	199
	0.2	51	100	154	201
	0	49	100	151	201
Average	-0.04	48.86	99.71	151.71	199.71

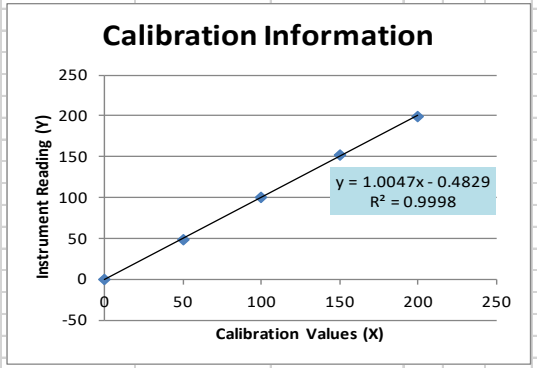
Best Fit Concentration	**49.75**	**100.47**	**150.23**	**200.47**
Point Difference (Best fit - Average)	**0.90**	**0.76**	**1.49**	**0.75**
Percent Difference (Best fit Conc vs. Avg Y values)	-1.80%	-0.76%	0.99%	-0.37%

r	slope (m_i)	intcpt (I_i)	lin reg
0.9999	1.0047	-0.4829	1.002

Can determine if calibration data meet Validation Template Guidance

- 1.5 ppb or
- 2% of best fit line

X	Y
0	-0.04
50	48.86
100	99.71
150	151.71
200	199.71



AQS 504 Data Assessment Tool

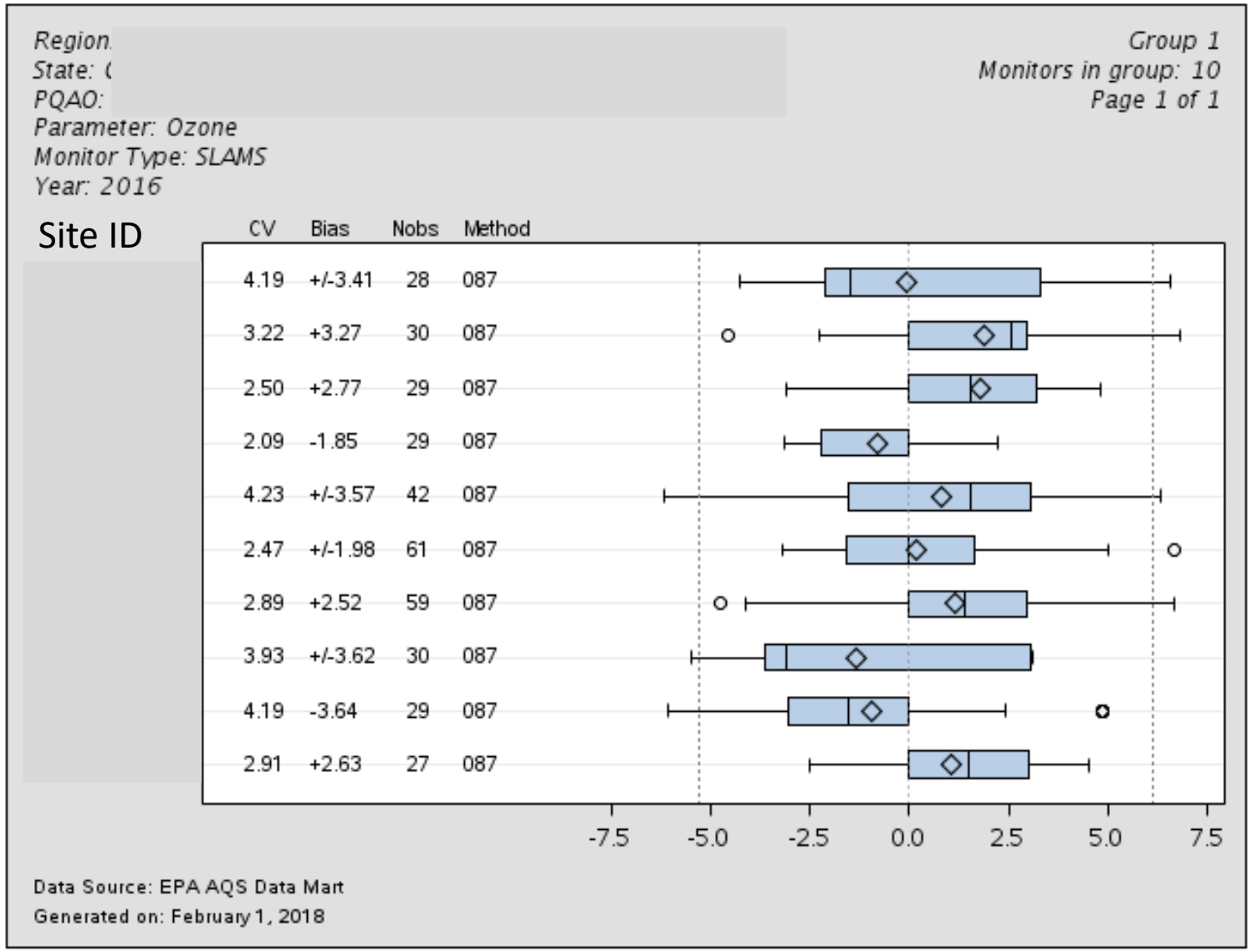
- Developed by EPA Region 3
- Run AMP504 Report for data you want to evaluate. Save as TXT
- Open 504 Assessment Excel Report (on AMTIC). It automatically opens a screen so you can find the txt file you want to import. Find file click open
- Converts 504 text file to an excel file and saves it as a separate file.
- Adds worksheets and additional information to the file.
- Sorts through data and identifies exceedances based on the criteria from Validation Templates
- Creates a final report

504 Report-Summary of Data Exceeding Acceptance Criteria

Assessment Type	County/ City Name	AQS ID	Parameter Code	Monitor Method Code	Monitor Method	Assessment Date
1-Point QC	**Any Town**	**xx-xxx-1013-1**	42401	560	Thermo Electron 43c-TLE/43i-TLE	**01/08/15**
1-Point QC	**Any Town**	**xx-xxx-1013-1**	42401	560	Thermo Electron 43c-TLE/43i-TLE	**01/22/15**
1-Point QC	**Any Town**	**xx-xxx-1013-1**	44201	47	THERMO ELECTRON 49	**01/08/15**
1-Point QC	**Any Town**	**xx-xxx-1013-1**	44201	47	THERMO ELECTRON 49	**06/04/15**

Monitor Conc.	Assessment Conc.	% Difference	Part 58 Appendix A Criteria	Last Valid Assessment Date	Last Valid % Difference	Number of Days Affected
8.9	8	11.3%	10.0%	**NONE**	**-1.1%**	**7**
9	8	12.5%	10.0%	**NONE**	**-1.0%**	**21**
0.043	0.04	7.5%	7.0%	**NONE**	**0.0%**	**7**
0.037	0.04	-7.5%	7.0%	**05/28/15**	**0.0%**	**7**

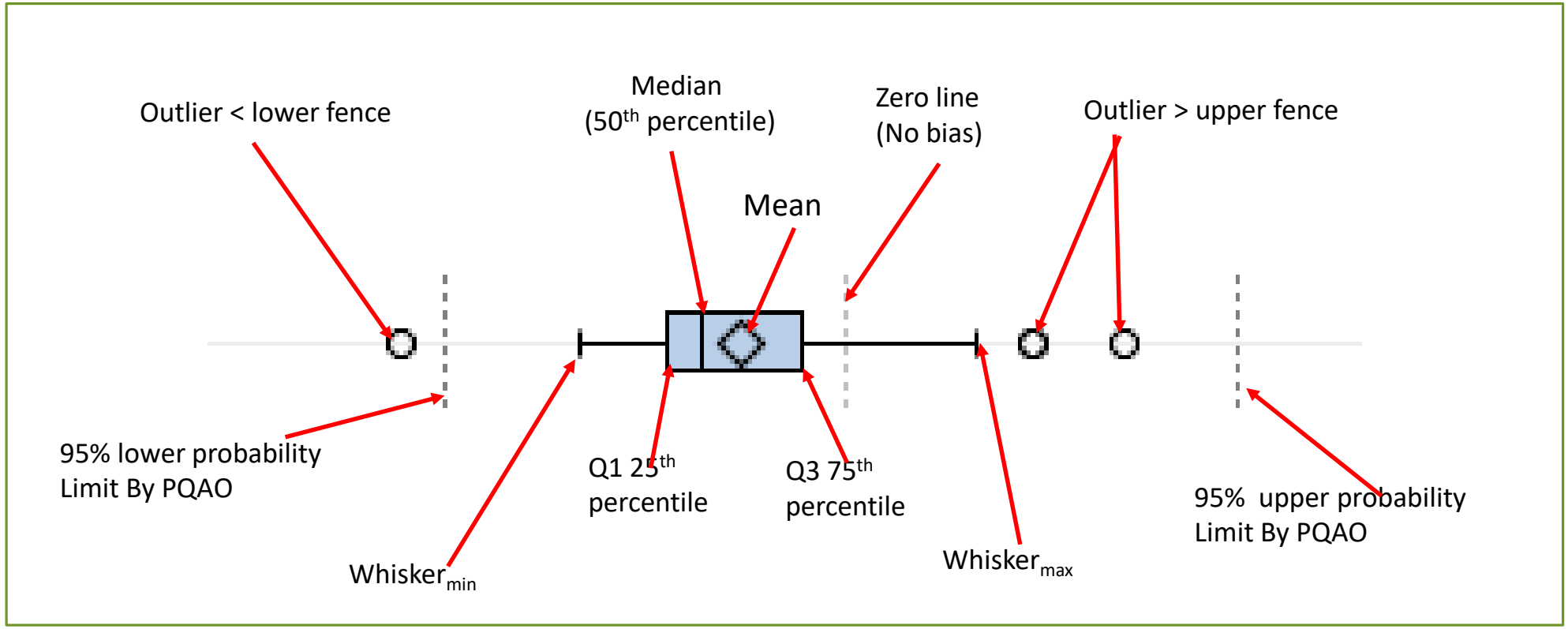
Single Point Precision and Bias Reports -Air Data



Selection Criteria

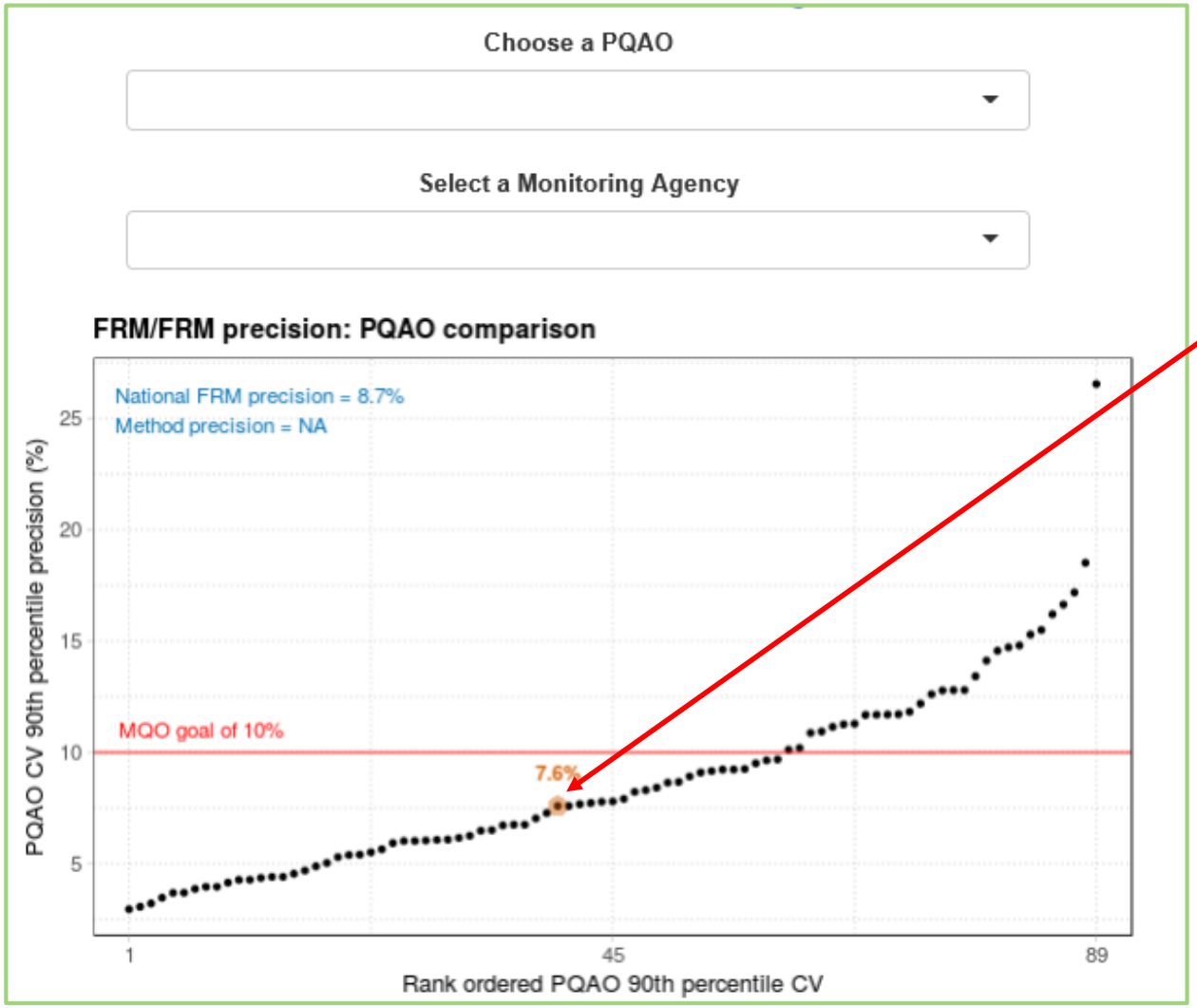
1. Pollutant
   - Ozone
2. Year
   - 2016
3. Domain
   - Select an EPA Region ...
   - -- or --
   - Select a State ...
   - -- or --
   - Select a PQAO ...
4. Bounds for graph
   (leave defaults to plot the full range of data)
   - Lower Default | Upper Default

Air Data-Single Point Precision and Bias Report Box and Whiskers Defined



Whisker$_{min}$ & Whisker$_{max}$-The lowest and highest values respectively that are found within the upper and lower fence. The upper and lower fences are defined as values between Q1 - (1.5*IQR) and Q3 + (1.5 * IQR), where "IQR" = the difference between Q3 and Q1.

R-Shiny Automated $PM_{2.5}$ FRM Data Quality Assessment



Three major features of this application which add value over existing assessments:

1. Visualizes the data
2. Compares the PQAO of interest to all other PQAO's as well as the MQO
3. Openly available (i.e., not password protected)

- Assessments available at PQAO level and monitoring agency level
- Includes up to 3 years of data for four indicators of $PM_{2.5}$ data quality:
  - **Collocated precision (shown)**
  - Bias via Performance Evaluation Program
  - Flow Rate Audits/Verifications
  - Field Blanks

https://stl-r-shiny.shinyapps.io/QVA_Dashboard/

R-Shiny Automated $PM_{2.5}$ FRM Data Quality Assessment

Right side graphics provide additional detail on QC data related to PQAO highlighted on the left

AQS API – web query service for AQS data

Data Available – useful for custom data assessments
Daily Summary
Annual Summary
QA Blanks
QA Collocated Assessments
QA Flow Rate Verifications
QA Flow Rate Audits
QA One Point QC Raw Data
QA PEP Audits

Example: returns collocated assessment data for FRM PM2.5 January 2013 where the PQAO is the Alabama Department of Environmental Management (agency 0013): https://aqs.epa.gov/data/api/qaCollocatedAssessments/byPQAO?email=test@aqs.api&key=test&param=88101&bdate=20130101&edate=20130131&pqao=0013

Live Demos of Data Assessment Tools

AQS AMP 504 CONTROL CHART; DASC; R-SHINY PM2.5 DATA ASSESSMENT TOOL; AQS API;

Summary

- Review QC data as it's generated…keep up with the small stuff and you'll avoid the big stuff
- Develop as many assessment tools as you can in data loggers and central office
- Graph/visualize the data. More may pop out at you.
  - Look at QA data across your sites. Maybe some operators are deviating from SOPs, maybe equipment at one or two sites is still passing but on the brink.
  - Look at data across years. Is there any overall trend? Maybe equipment is just getting old
- Take advantage of the AQS reports and tools available

Questions?