# DEVELOPMENT OF MORE STATISTICALLY ROBUST MEASUREMENT QUALITY OBJECTIVES (MQOs) FOR LABORATORY BIAS FOR THE NATTS NETWORK

Battelle Memorial Institute
505 King Avenue
Columbus, OH  43201

Prepared for:

Margaret Dougherty, Task Order Project Officer
Dennis Mikel, US EPA Alternate Task Order Project Officer
U.S. Environmental Protection Agency
Research Triangle Park, NC 27709

Task B
Task Order # EP-G11D-00028
Contract No. GS-10F-0275K

July 17, 2012

## Battelle Disclaimer

This report is a work prepared for the United States Government by Battelle. In no event shall either the United States Government or Battelle have any responsibility or liability for any consequences of any use, misuse, inability to use, or reliance upon the information contained herein, nor does either warrant or otherwise represent in any way the accuracy, adequacy, efficacy, or applicability of the contents hereof.

## EPA Disclaimer

The material in this document has not been subject to Agency technical and policy review. Views expressed by the authors are their own and do not necessarily reflect those of the U.S. Environmental Protection Agency. Mention of trade names, products, or services does not convey, and should not be interpreted as conveying, official EPA approval, endorsement, or recommendation.

# Table of Contents

Page

## Executive Summary

A series of statistical analyses were performed using historical proficiency testing (PT) data collected from 2004 to 2010 for the National Air Toxics Trends Stations (NATTS) ambient air monitoring program to (1) assess the attainment of the current measurement quality objective (MQO), i.e. acceptance criterion, for laboratory bias, and to (2) develop more robust, compound-specific MQOs that broadly account for the capabilities of participating laboratories as well as for differences in the various analytical methods used to quantify the different classes of compounds measured in the NATTS network. The current NATTS bias MQO states that a participating laboratory must be capable of reporting a PT sample concentration that is within 25% of the sample's stated ("true") concentration. To develop more robust, compound-specific MQOs, a "variance-based" approach was adopted that accounts for observed variability in the data within batches of PT samples (e.g., 2 times the pooled standard deviation, $\sigma_{pooled}$). In addition, comparisons were performed to investigate differences in laboratory performance depending on whether the target value for a given batch of PT samples was assigned as the average result from all participating laboratories (   ) or as a stated "true" value.

Two important recommendations resulting from the analyses present in this report are as follows:

- The target value against which participating laboratories results are compared should be the batch-specific average across labs (   ) rather than a "true" value assigned as either the theoretical spiked amount or the value obtained from confirmatory analyses.

- Using $\pm 2\sigma_{pooled}$ as the lab bias MQO appears to be the most reliable method of those examined to ensure that the laboratories will be able to meet the MQO more consistently since such an MQO accounts for the demonstrated capabilities of the analytical methods as they are applied across many laboratories over time.

Variance-based MQOs for laboratory bias were calculated as $2\sigma_{pooled}$ for six analytes of particular interest in the NATTS network. The proposed laboratory bias MQO for each of those analytes is shown in the table below. As can be seen in the table, the current bias requirement of < 25% may be reasonable for arsenic; however, it may be too generous for formaldehyde, yet too stringent for other analytes, especially acrolein.

**Proposed laboratory bias MQOs based on historical NATTS PT data**

| Analyte | Proposed Lab Bias MQO |
|---|---|
| Acrolein | 60 to 70% |
| Arsenic | 25 to 30% |
| Benzene | 30 to 35% |
| 1,3-butadiene | 35% |
| Formaldehyde | 15% |
| Nickel | 30 to 40% |

## 1. Introduction

Since 2003, the US Environmental Protection Agency (EPA) has operated the National Air Toxics Trends Stations (NATTS) ambient air monitoring program, which generates long-term, quality assured, standardized ambient air toxics data that can be used in the following activities:

- establishing trends and evaluating the effectiveness of EPA's air toxics emissions reduction strategies;
- characterizing ambient concentrations/deposition of priority air toxics in local areas;
- supporting, evaluating, and improving air quality models, and as input to source-receptor models; and
- supporting scientific studies to better understand the relationship between ambient air toxics concentrations, human exposure, and health effects from these exposures.

For four toxic compounds that serve as high-risk drivers (the volatile organic compounds [VOCs] benzene and 1,3-butadiene; $PM_{10}$ arsenic; and formaldehyde, a carbonyl compound), EPA assesses data collected from the NATTS program in order to determine whether the participating laboratories are achieving the specified measurement quality objectives (MQOs), also known as acceptance criteria, for the data quality indicators (DQIs) of completeness, precision, laboratory bias (accuracy), and sensitivity (method detection limits). EPA established the current MQOs for the various DQIs at the start of the program. Recently, EPA has indicated a need to determine whether to recommend modifications to the MQOs, and in particular, whether the statistical robustness of the MQOs can be improved upon such that the MQOs better reflect the demonstrated capabilities of the agencies comprising the NATTS program.

Under Task B of Task Order # EP-G11D-00028 (Contract No. GS-10F-0275K), Battelle performed statistical analyses on historical proficiency testing (PT) data from the NATTS program to provide information for EPA to reassess the NATTS' current MQO for laboratory bias. This MQO states that, through the analysis of PT samples conducted on a quarterly basis, a participating laboratory must be capable of reporting a sample concentration that is within 25% of the sample's stated ("true") concentration, T, where T is determined by the PT provider laboratory either as a theoretical spike amount or from the analysis of confirmatory samples. Battelle designed its analyses to answer questions such as:

- To what extent are labs meeting the current < 25% laboratory bias MQO (e.g., percent difference of the reported measurement to T is within 25%)?
- Should the MQO be modified
  - to be analyte-specific?
  - from the current fixed percentage to a "variance-based" approach that accounts for observed variability in the data within batches of PT samples (i.e., 2 times the batch-specific or pooled standard deviation $\sigma$)?
- Is the means for determining the target value used to assess the MQO appropriate, or should it change (e.g., from a stated "true" value, T, to the batch-specific average across labs)?

While it was also of interest to perform data analyses to investigate whether more statistically robust acceptance criteria could be established for the confirmatory analyses that a PT provider laboratory performs on each batch of PT samples that it creates (i.e., to confirm that the PT samples were acceptably prepared and spiked PT amounts acceptably recovered), no relevant PT data could be obtained to address this analysis at this time.

The analyses presented in this report are divided into two phases:

- Phase I assessed laboratory bias relative to percent difference from the “true” value T, including a comparison of the use of T and the average across labs as measures of the target value for PT samples.
- Phase II investigated the establishment of a variance-based MQO for laboratory bias.

## 1.1 Data Description

The statistical analysis utilized historical NATTS PT sample data which EPA provided to Battelle on 11/9/2011. These PT sample data were collected from participating labs over the period from 2004 to 2010. They represent 37 different analytes (15 VOCs, 10 particle-bound metals species [metals], 5 carbonyls, and 8 polycyclic aromatic hydrocarbons [PAHs]) which are listed in Table 1. The following observations were made from the data:

- Data were available for only one or two years, and had not been collected since 2008 or earlier, for antimony, chromium, mercury, acetone, and crotonaldehyde.
- PT data for the PAHs were available only from 2009 to 2010.
- PT data for hexavalent chromium were available for only three labs and only in 2010.

**Table 1. Analytes Contained in the Historical NATTS PT Sample Database**

| VOCs | Metals | Carbonyls | PAHs |
|---|---|---|---|
| ***1,1,2,2-tetrachloroethane*** | Antimony | ***Acetaldehyde*** | Acenaphthene |
| ***1,2-dibromoethane*** | ***Arsenic*** | Acetone | Anthracene |
| ***1,2-dichloroethane*** | ***Beryllium*** | Crotonaldehyde | Benzo(a)pyrene |
| ***1,2-dichloropropane*** | ***Cadmium*** | ***Formaldehyde*** | Fluoranthene |
| ***1,3-butadiene*** | Chromium | | Fluorene |
| ***Acrolein**** | ***Lead*** | | Naphthalene |
| ***Benzene*** | ***Manganese*** | | Phenanthrene |
| ***Carbon tetrachloride*** | Mercury | | Pyrene |
| ***Chloroform*** | ***Nickel*** | | |
| ***cis-1,3-dichloropropene*** | Hexavalent Chromium | | |
| ***Dichloromethane*** | | | |
| ***Tetrachloroethylene (PERC)*** | | | |
| ***trans-1,3-dichloropropene*** | | | |
| ***Trichloroethylene*** | | | |
| ***Vinyl chloride*** | | | |

Note: Analytes in bold italics indicate those included in the Phase I analyses
* Classified as a carbonyl in 2004, but as a VOC thereafter.

For the remaining 23 analytes listed in bold italics, PT data were available for three or more years, and thus were the focus of the analyses under Phase I. Phase II focused on six analytes of stated high importance to EPA: acrolein, arsenic, benzene, butadiene, formaldehyde, and nickel.

All historical NATTS data from 2004-2010 were used in the analyses in this report with the exception of data for some batches in 2004 and 2005 for which the "true" value T far exceeds the PT spiking levels of the current NATTS program. Data were excluded for VOCs with T greater than 10 ppb, metals with T greater than 10 μg/filter, and carbonyls with T greater than 10 μg/cartridge. Data were also excluded for batches of beryllium and nickel with T greater than 8 μg/filter, and batches of cis-1,3-dichloropropene with T greater than 8 ppb.

For each analyte, this report presents bar charts of the percent difference from some representation of the target value of the spiked amount or concentration for the PT samples, along with the results of an analysis of variance to investigate whether variability tends to be higher in certain situations (e.g., for specific labs, at the inception of the NATTS program versus later years, and/or for other factors). For example, labs identified as having higher variability (perhaps due to use of different analytical methods) may need to be handled differently or perhaps excluded from the summary calculations.

## 2. Phase I Analysis

### 2.1 Bar Charts

The bar charts presented in Appendix A illustrate the typical distribution of measurements that characterize laboratory bias. The skewness and dispersion of the results displayed in these bar charts help assess the extent to which variability in laboratory bias is present across samples and laboratories. Two bar charts were prepared for each of the 23 analytes (listed in bold italics in Table 1) for which historic PT data were collected and available in sufficient quantities across years. Each bar chart summarizes percent difference in the reported concentrations relative to a specified target value that represents the actual concentration of the analyte in the samples. For a given analyte, the horizontal and vertical axes for both bar charts are on the same scale.

***Bar Chart #1*** *(top chart on each page of Appendix A)****.*** For a given analyte, the first of the two bar charts presents percent difference with respect to the "true" value T which was presumed to be known and constant for all PT samples in a given batch:

$$\text{Percent difference} = PD_1 = 100 * (\text{Result-T})/\text{T}$$

A "batch" represents the PT samples provided to all laboratories in a given quarter and containing a specified concentration, T, of the analyte in question. For some analytes such as butadiene and benzene (and sometimes acrolein), T typically corresponded to the average of one or more confirmatory analyses performed by a single laboratory. For other analytes such as arsenic and formaldehyde, T is a theoretical amount or concentration that was calculated by the PT provider laboratory. Because the value of T was assumed to be constant for all samples in a batch, this percent difference calculation can be used to assess the current MQO. Note that any sample measurement equal to T would have a percent difference of 0%. Red vertical lines on the bar charts represent $PD_1$ values of ± 25%. Therefore, observations falling outside of these red vertical lines in the bar chart do not meet EPA's current MQO for laboratory bias.

Computing a Z-score based on T and the standard deviation ($\sigma$) of the measured concentrations across all labs in a given batch represents an accepted best practice[1] for assessing laboratory precision:

$$Z = (\text{Result} - T)/\sigma$$

Generally if $-2 \leq Z \leq 2$, the Z-scores can be assumed to follow a standard normal distribution, and T is presumed to be the actual known true value of the PT sample, the laboratory is said to meet the requirements for laboratory bias.[2] Another way to state this is that if the result is more than two standard deviations from the mean (where T represents the mean), the result is unacceptable.

Because T and $\sigma$ vary with each batch, it is not possible to draw lines on the bar chart to represent distances of $\pm 2\sigma$ from T since each chart contains multiple batches of data. Instead, those observations that fall outside of $2\sigma$ from T for a given batch are indicated in red on the bar chart.

The set of PT sample measurements may include statistical outliers (e.g., a measurement that exceeds T by a large amount, such as 300%) that will have a large effect on statistics such as the batch-specific average and standard deviation across laboratories (  and $\sigma$, respectively). For this reason, an outlier test (Grubbs test at a 1% significance level) was performed on the measurements within each batch to identify the presence of outliers. Values for   and $\sigma$ were then computed for each batch after excluding any outliers.

Finally, as part of Phase II, a pooled standard deviation ($\sigma_{pooled}$) in the PT sample measurements across all batches was estimated with an analysis of variance method. Because $\sigma_{pooled}$ is calculated across batches, it is constant for a given analyte; values of $T \pm 2\sigma_{pooled}$ are presented by vertical green lines within these bar charts. The calculation of $\sigma_{pooled}$ is discussed in more detail in the Phase II section.

Note that distributions of reported percent differences for several analytes presented in Appendix A are not well centered around T. For example, more than 2/3 of the measurements for tetrachloroethylene (PERC) are less than T which could possibly indicate that samples tend to degrade between the time of spike and analysis. On the other hand, more than 2/3 of the measurements for beryllium are greater than T. One possible explanation for such behavior is that T is a theoretical spiked amount, and the samples were consistently spiked higher than the targeted level or the media were persistently contamined. Another observation is that many of

[1] For example, see ASTM E691-11, "Standard practice for conducting an interlaboratory study to determine the precision of a test method." Available at http://webstore.ansi.org/RecordDetail.aspx?sku=ASTM+E691-11.

[2] The implicit assumption of ASTM E691 is that T is unknown which limits the scope of the practice to the assessment of precision. Bias may be evaluated by assuming that T is known.

the histograms have a heavy upper tail. Such distributions are consistent with the fact that there is no upper bound on positive percent differences, but the lowest observable percent difference is -100% (if the reported measurement is 0).

***Bar Chart #2*** *(bottom chart on each page of Appendix A)****.*** The second of the two bar charts for each analyte plots an alternative calculation of percent difference:

$$\text{Alternative percent difference} = PD_2 = 100 * (\text{Result-} \ \ )/$$

Note that for each analyte, bar chart #2 differs from bar chart #1 only in the use of $PD_2$ rather than $PD_1$ (i.e., replacement of T with the batch-specific mean   as a representation of the target concentration in the sample). Compared to T, the value   may be a better estimate of the actual concentration applied to a batch of PT samples compared to T, as the former is either a theoretical amount or a value based on one or two analyses by a single (only presumably superior) laboratory whereas the latter is an average across numerous labs and likely better represents the population of all PT samples prepared in a given batch.

In these bar charts, red vertical lines represent $PD_2$ values of ± 25%, while vertical green lines indicate when $PD_2$ equals $\pm 2\sigma_{pooled}$. Also, the Z-score introduced earlier is now defined as

$$Z = (\text{Result-} \ \ )/ \ \sigma$$

As before, it is not possible to draw lines on the bar chart to represent distances of ± 2σ from   because   and σ vary with each batch. Thus, those observations that fall outside of 2σ from   for a given batch are indicated in red on the bar chart. Note that if |Z| < 2 implies that a laboratory meets the requirements for laboratory bias, then a laboratory meets the bias requirement if |Result-  | < 2σ, or $|PD_2| < 200*\sigma/$ . Therefore, a redefined MQO based on Z-score can still be formulated to include a percent difference, $PD_2$.

For each of the 23 analytes, both charts include a text box indicating the percentage of observations, including outliers, that fall outside of ± 25%, outside of ± 2σ and outside of ± $2\sigma_{pooled}$, including outliers. These percentages are also listed in Table 2. They correspond to:

- Bar chart #1: T ± 25% (the current MQO – shaded column), T ± 2σ, T ± $2\sigma_{pooled}$,
- Bar chart #2:   ± 25%,   ± 2σ,   ± $2\sigma_{pooled}$.

Table 2 also shows the percentage of samples that were identified as outliers by Grubb's test with a significance level of 1%. Note that the percentage of observations outside of   ± 2σ is greater than the expected 5% (approximately 95% of normally distributed data can be expected to be between   ± 2σ). This is because of deviation from normality and the presence of outliers. If outliers are excluded, the number of observations outside of 2σ is about what would be expected assuming a normal distribution. Therefore to assess the stringency of the current MQO of ± 25% the percentage of reported results outside of   ± 25% may be compared to the percentage of results outside of   ± 2σ. More observations are outside of   ± 25% than   ± 2σ for 20 analytes, and assuming that a reasonable goal would be that approximately 95% of participating labs perform acceptably well, the current ± 25% MQO may be too difficult for labs to consistently attain for many compounds. However, labs are measuring acetaldehyde and formaldehyde more accurately than ± 25%, consistent with the conclusion that the TO-11a analysis methodology for carbonyls has a lower inherent bias than other methods and, as a result, the current lab bias MQO for these two compounds may be too generous.

**Table 2. Percentage of Historic PT Measurements that are Outside of Specified Targets**

| | | | | Bar Chart #1 | | | Bar Chart #2 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Analyte** | **# of Labor-atories** | **# of Obser-vations** | **% of Outliers** | **% Outside T ± 25%** | **% Outside T ± 2σ** | **% Outside $T \pm 2\sigma_{pooled}$** | **% Outside ± 25%** | **% Outside ± 2σ** | **% Outside $\pm 2\sigma_{pooled}$** |
| 1,1,2,2-tetrachloroethane | 34 | 313 | 2.2% | 12.5% | 8.9% | 7.7% | 11.2% | 7.0% | 8.3% |
| 1,2-dibromoethane | 35 | 310 | 2.9% | 11.3% | 9.0% | 8.7% | 10.3% | 7.7% | 8.7% |
| 1,2-dichloroethane | 35 | 335 | 3.0% | 8.1% | 11.0% | 10.5% | 6.9% | 6.3% | 8.1% |
| 1,2-dichloropropane | 33 | 286 | 2.8% | 12.6% | 15.4% | 13.6% | 7.7% | 7.0% | 7.7% |
| 1,3-Butadiene | 34 | 321 | 2.8% | 16.5% | 10.3% | 10.0% | 15.3% | 9.0% | 8.4% |
| Acetaldehyde | 30 | 242 | 1.7% | 4.5% | 12.8% | 12.0% | 3.7% | 9.1% | 7.9% |
| Acrolein | 25 | 152 | 1.3% | 40.1% | 15.1% | 9.2% | 39.5% | 7.9% | 7.2% |
| Arsenic | 25 | 225 | 2.2% | 12.9% | 12.9% | 5.8% | 8.9% | 5.8% | 8.0% |
| Benzene | 35 | 300 | 1.0% | 9.0% | 10.3% | 4.3% | 7.7% | 5.3% | 4.0% |
| Beryllium | 25 | 223 | 1.8% | 18.8% | 14.8% | 11.7% | 10.8% | 6.7% | 9.4% |
| Cadmium | 25 | 229 | 1.3% | 6.6% | 10.0% | 4.4% | 7.0% | 6.1% | 5.7% |
| Carbon tetrachloride | 35 | 332 | 1.5% | 11.7% | 11.4% | 8.4% | 10.2% | 6.9% | 7.2% |
| Chloroform | 35 | 309 | 2.3% | 10.4% | 16.2% | 15.2% | 6.8% | 6.1% | 7.4% |
| cis-1,3-dichloropropene | 34 | 301 | 1.7% | 15.3% | 9.0% | 9.3% | 11.6% | 7.3% | 7.6% |
| Dichloromethane | 35 | 319 | 1.9% | 11.6% | 8.8% | 7.2% | 11.0% | 8.5% | 6.6% |
| Formaldehyde | 30 | 242 | 1.7% | 2.9% | 21.9% | 16.1% | 2.5% | 6.6% | 7.4% |
| Lead | 25 | 228 | 5.3% | 12.3% | 11.4% | 13.2% | 7.9% | 8.8% | 8.8% |
| Manganese | 25 | 228 | 2.2% | 15.4% | 18.4% | 12.7% | 10.5% | 7.5% | 7.9% |
| Nickel | 25 | 226 | 2.7% | 14.2% | 10.6% | 8.0% | 13.3% | 7.5% | 7.5% |
| Tetrachloroethylene (PERC) | 35 | 307 | 1.0% | 14.3% | 11.4% | 7.5% | 9.4% | 3.9% | 4.6% |
| trans-1,3-dichloropropene | 34 | 301 | 3.7% | 18.6% | 12.0% | 8.3% | 13.6% | 9.0% | 8.3% |
| Trichloroethylene | 35 | 333 | 1.5% | 8.4% | 9.0% | 5.1% | 7.2% | 6.0% | 5.1% |
| Vinyl chloride | 35 | 307 | 2.9% | 13.7% | 15.6% | 10.4% | 12.1% | 7.5% | 8.1% |

Note: The current MQO is represented in the shaded column.

The percentage outside of T ± 25% is greater than ± 25% for all but one analyte, cadmium (where the values are very similar). This is expected, as the aggregate of all observed values are used to calculate , thereby leading to a better estimate of the actual spiked amount. T will be somewhat different from at least due to random variation, but also because the single lab measuring T may use a more accurate measurement technique than the other labs (for instance, in the case of the use of gas chromatography with flame ionization detection for confirmation of VOCs concentrations in 2008-2010) or because of error in the estimation of the spiked level (if T represents the spiked level).

A compromise between the approach of using a standard MQO of ± 25% and the approach of comparing to ± 2σ for each spiked level would be to use $\sigma_{pooled}$ to estimate a reasonable range for the MQO. For example the range ± $2\sigma_{pooled}$ covers more samples than ± 25% for many of the analytes, with the exception of formaldehyde and acetaldehyde. This approach is discussed in more detail in the Phase II section.

## 2.2 Detailed Phase I Analyses (in preparation for Phase II work)

### 2.2.1 Identifying Labs with High Variability in PT Measurements

Additional analyses were performed on the PT data for the 23 analytes to investigate whether variability within a batch (as measured by σ) tends to be higher in certain situations. For example, it is important to identify labs whose results tend to be much more variable than other labs, perhaps due to analytical method, potential difficulties in method implementation, or other factors. It is also possible that the overall variability of the measurements for a given analyte across all labs may have been greater at the inception of the NATTS program compared to performance in later years.

To screen for labs with unusually high variability, the proportion of outliers was examined by use of the variable “LabCode” in the PT dataset. “LabCode” is unique for each lab and analyte category (VOC, carbonyl, or metal). In this way results of a particular lab’s VOC measurements are analyzed separately from the same lab’s metals measurements. Because the Grubb’s test for outliers was performed at a 1% significance level, there is a 1% chance that each sample is flagged as an outlier at random. For each lab code, the probability that the number of observed outliers or more would occur at random from the total number of NATTS samples submitted by that LabCode was computed using the following formula:

where: $x$ = the number of observed outliers
$n$ = the total number of samples
$\alpha$ = the significance level of the Grubb’s test (1%)

Labs with a value of $p$ less than 1% are listed in Table 3. The analyses in Phase II were performed with and without the data for these labs.

**Table 3. LabCodes Associated With an Unusually High Number of Outliers**

| LabCode[a] | Lab Name | Number of Outliers | Number of PT Samples | p |
|---|---|---|---|---|
| 02-01-V | Rochester & Bronx, NY | 12 | 221 | <0.0001 |
| 04-02-V | Chesterfield, SC | 21 | 223 | <0.0001 |
| 04-04-M | Atlanta, GA | 10 | 88 | <0.0001 |
| 05-06-M | Indiana | 6 | 30 | <0.0001 |
| 05-07-V | Ohio EPA DAPC | 12 | 77 | <0.0001 |
| 07-02-V | Univ. of Iowa | 6 | 169 | 0.0074 |
| 08-03-V | EPA – NEIC | 11 | 28 | <0.0001 |
| 10-01-M | Seattle, WA | 6 | 54 | <0.0001 |
| 10-01-V | Seattle , WA | 5 | 77 | 0.0011 |
| 10-01B-V | R.J. Lee Group | 9 | 118 | <0.0001 |

[a] V = VOC; M = metals

### 2.2.2 Increase in Precision over Time

Statistical analyses were also performed on the six analytes that were the focus of Phase II (acrolein, arsenic, benzene, butadiene, formaldehyde, and nickel) to determine if the precision of the measurements for a given analyte across all labs had changed over time. Specifically the concern is that the variability may have been greater at the inception of the NATTS program versus later years and that the older measurements will artificially increase the observed measurement variability. A Brown and Forsythe equality of variance test was performed on the alternate percent difference values defined above ($PD_2$). The data were grouped into years, so that a group consisted of all of the $PD_2$ measurements from all labs for a given analyte in a given year. First, a Brown and Forsythe equality of variance test was performed to determine if there were any statistically significant differences between the variances of the groups.

Table 4 shows the four analytes that showed statistical evidence of some difference between groups. Plots of the standard deviation of the percent difference by quarter for these analytes are shown in Figures 1 through 4. It is apparent from the plots that there was a much larger variability in percent difference for arsenic and nickel in 2009. In fact, this phenomenon can be seen for most of the metals in 2009 (see section 2.2.3); as such, Phase II analyses for these two metals were performed with and without their 2009 data. The standard deviation for benzene was very high one quarter in 2006, but is relatively consistent otherwise. However, since there is no additional justification for the removal of data for this one quarter, it was included in all analyses. Acrolein is the only analyte that shows evidence of a decrease in variability since the beginning of the NATTS program. For acrolein, a series of sequential paired tests (with adequate multiple comparison adjustment to the p-value) of equality of variance was performed. The following comparisons were made:

- 2006 vs. 2007-2010
- 2006-2007 vs. 2008-2010
- 2006-2008 vs. 2009-2010
- 2006-2009 vs. 2010

The comparison of 2006 to 2007-2010 was statistically significant (p=0.0024), which suggests that the variability in 2006 was statistically higher than the variability in subsequent years. As a result, the Phase II analyses were performed with and without the 2006 data for acrolein.

**Table 4. Analytes with Statistical Evidence of Inequality of Variance of Percent Difference across Years**

| Analyte | p-value |
|---|---|
| Acrolein | 0.0142 |
| Arsenic | <0.0001 |
| Benzene | 0.0422 |
| Nickel | <0.0001 |

### 2.2.3 Metals in 2009

Because an unusually high standard deviation of the percent difference from the mean was observed for arsenic and nickel, the data for the other metals (beryllium, cadmium, lead and manganese) were plotted. These plots are located in Appendix B. A similar spike in variability is apparent in these plots as well. For this reason, Phase II analyses were performed with and without the 2009 data for arsenic and nickel.

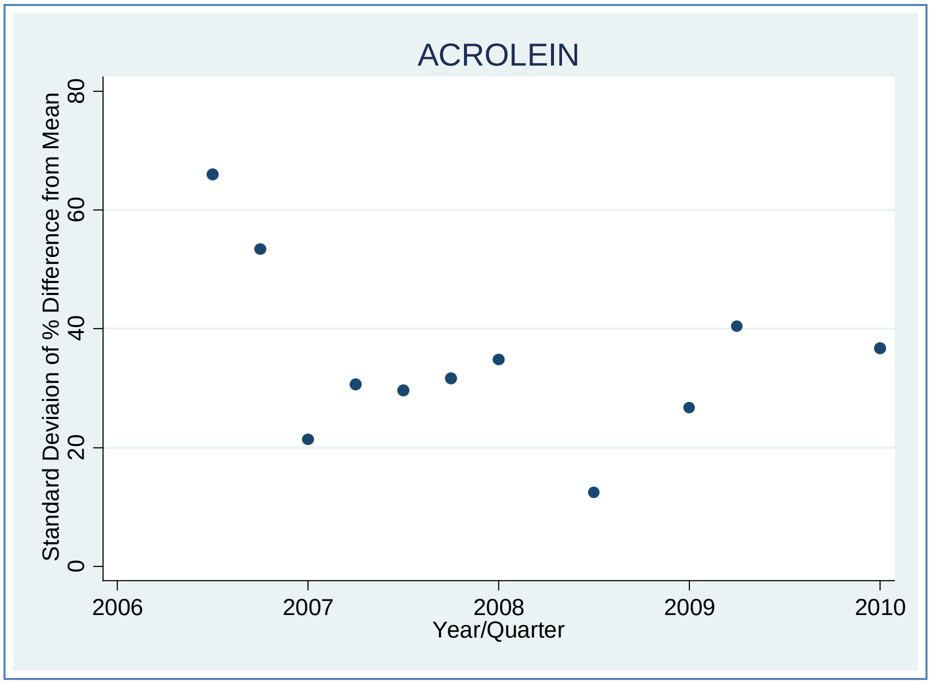


**Figure 1. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Acrolein**

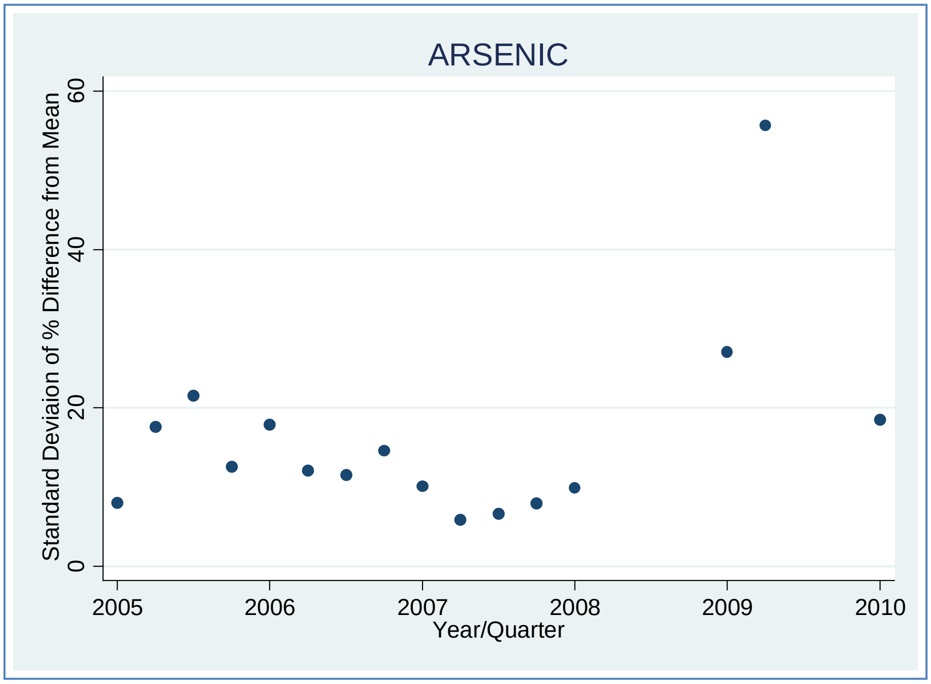


**Figure 2. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Arsenic**

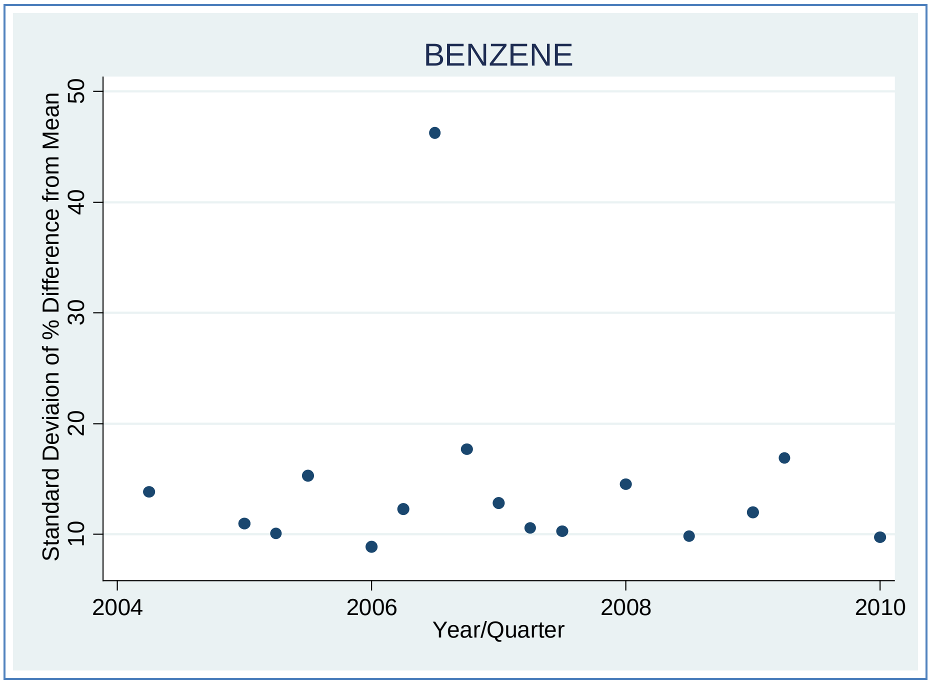


**Figure 3. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Benzene**

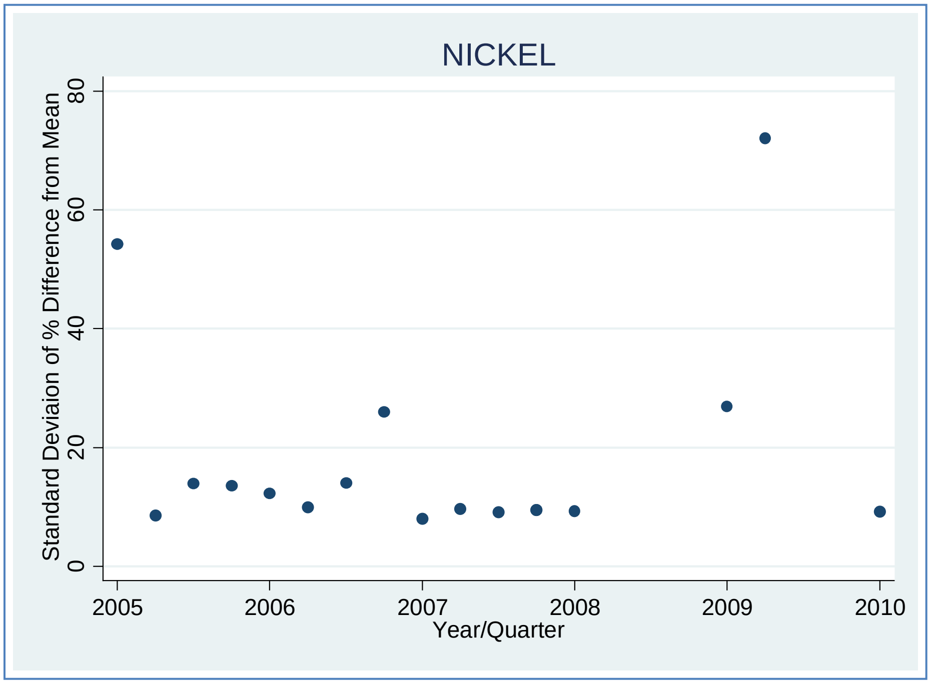


**Figure 4. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Nickel**

### 2.3 Phase I Discussion and Conclusions

For each analyte, the first of the two bar charts in Appendix A shows the relationship of the historical PT samples from all labs to the current MQO of T ± 25%. More than 5% of the observations do not meet the current MQO for 21 of the 23 analytes; for acrolein, more than 40% of the PT samples do not meet the MQO. These results imply that a large percentage of the laboratories are likely not performing to EPA's expectations. The most logical explanation is that the analytical methods and procedures as implemented at most of the labs are incapable of consistently meeting the MQO. The extent to which the assigned target value varies from the actual sample concentration for a given batch of PT samples can also contribute to the laboratory bias assessment. If a single lab is used to determine T (based on one or more measurements), and this lab does not use a more accurate analytical technique than the other labs, then using an average measurement across all participating labs to characterize the target value would yield a more accurate estimate of the actual spiked value. Comparing the percentage of samples outside of T ± 25% to the percentage of samples outside of ± 25% demonstrates the effect of using the lab average instead of T: across all analytes, approximately 2.5% more samples would meet the MQO if the lab average rather than T were used to estimate the target value. Even so, for nearly half of the analytes (12 of 23), more than 10% of the observations fall outside of ± 25%, indicating that even with this change, a fair amount of labs would still not meet such an MQO consistently. A MQO based on the historical variability of the measurements would align the lab bias requirement with the overall capability of the labs. Nine of the 23 analytes have more than 10% of the observations outside of T ± $2\sigma_{pooled}$, while no analyte has more than 10% of the observations outside of ± $2\sigma$ or ± $2\sigma_{pooled}$.

## 3. Phase II Analysis

Phase II explores the establishment of a variance-based MQO for laboratory bias for six analytes: acrolein, arsenic, benzene, butadiene, formaldehyde, and nickel. The goal of this analysis was to calculate a single (pooled) estimate of the standard deviation of the percent difference across all labs and years for each analyte, excluding outliers as appropriate. Such estimates will serve as the basis for proposing alternative MQOs that are analyte-specific and account for the capabilities of participating laboratories as well as for differences in the various analytical methods used to quantify the different classes of compounds measured in the NATTS network.

### 3.1 Analysis of Variance Approach to Estimating Pooled Standard Deviation

With "batch" defined as a collection of samples for a given analyte with a unique year, quarter and target value, a random effects analysis of variance (ANOVA) model was fit for each analyte:

where: $y_{ij}$ is $PD_2$ for the $i^{th}$ batch and the $j^{th}$ lab,
$\mu$ is the mean $PD_2$across all labs and batches,
$\delta_i$ is effect of the $i^{th}$ batch, and
$\varepsilon_{ij}$ is the error term

The mean squared error of this model fit is the estimate of the pooled variance of the error term of this model, and therefore, the square root of this value is the pooled estimate of the within-batch standard deviation (i.e., $\sigma_{pooled}$). By modeling $PD_2$ rather than the sample measurement, $\sigma_{pooled}$ can be expressed as a percentage of the batch-specific lab average (similar to the current MQO of 25%).

For each of the six analytes, $\sigma_{pooled}$ was estimated under each of two conditions:

- A: The outliers identified in Phase I by Grubb's test performed at the 1% significance level were excluded.
- B: In addition to A,
    - the measurements for the 10 lab/analyte classes identified in Table 3 (Section 2.2.1) were excluded,
    - acrolein data from 2006 were excluded (due to a considerably high degree of variability compared to subsequent years).

For the two metals (arsenic and nickel), $\sigma_{pooled}$ was estimated under an additional condition:

- C: In addition to B, all data from 2009 were excluded

Values of $\sigma_{pooled}$ are listed in Table 5 under each of these three conditions. Multiplication of these pooled standard deviations by 2 yields an estimate of the new, more statistically based MQO, based on analysis of historical PT data**.**

**Table 5. Estimates of $\sigma_{pooled}$ and $2\sigma_{pooled}$ for Percent Difference Under Conditions A, B and C, for Six Analytes**

**$\sigma_{pooled}$**

| Analyte | Condition A | Condition B | Condition C |
|---|---|---|---|
| Acrolein | 35.6% | 29.4% | |
| Arsenic | 15.1% | 12.8% | 12.3% |
| Benzene | 16.6% | 14.2% | |
| 1,3-butadiene | 18.3% | 17.1% | |
| Formaldehyde | 7.6% | 7.6% | |
| Nickel | 18.9% | 16.2% | 15.3% |

**$2\sigma_{pooled}$ – estimate of new bias MQO based on historical PT data**

| Analyte | Condition A | Condition B | Condition C |
|---|---|---|---|
| Acrolein | 71% | 59% | |
| Arsenic | 30% | 26% | 25% |
| Benzene | 33% | 28% | |
| 1,3-butadiene | 36% | 34% | |
| Formaldehyde | 15% | 15% | |
| Nickel | 38% | 32% | 31% |

## 3.2 Scatterplots

For each combination of the six analytes and either two or three conditions for calculating $\sigma_{pooled}$ (A, B, C), Appendix C contains scatter plots of $PD_2$ versus the batch-specific lab average ( ). The black line in each plot indicates the lab averages, the red lines indicate the lab average ± 25% and the green lines indicate the lab average ± $2\sigma_{pooled}$. These figures demonstrate how laboratory performance compares to a MQO (lab average ± 25%) similar to the one used at present (T ± 25%) and to a more statistically robust MQO of lab average ± $2\sigma_{pooled}$, as well as the relative effect that the various data exclusions have on the value of $\sigma_{pooled}$.

Under the assumption that the target concentration T is represented by the lab average, the scatterplots show what to expect if the MQO of ± 25% were to be replaced with an MQO of $\pm 2\sigma_{pooled}$. For arsenic, Table 5 indicates that the MQO would be ± 30%, ± 26% or ± 25% depending on whether data from conditions A, B or C are used, respectively. (More simply, the MQO would be a value between ± 25-30%.) Similarly, the MQO would be between ± 30-40% for nickel, ± 30-35% for benzene, ± 35% for 1-3,butadiene, ± 15% for formaldehyde and between ± 60-70% for acrolein.

Table 6 shows the percentage of outliers (of all reported results) and the percentages of PT results outside of ± $2\sigma_{pooled}$ and outside of T ± $2\sigma_{pooled}$, calculated both with and without outliers (based on Condition A). After excluding outliers, the percentage of results outside of T

$\pm 2\sigma_{pooled}$ was as high as 14.4% (for formaldehyde); however the percentage of measurements for each of the six analytes outside of $\pm 2\sigma_{pooled}$ was about 5%. Using $\pm 2\sigma_{pooled}$ appears to be the most reliable method of those examined in this report to ensure that the labs will be able to meet the MQO more consistently.

**Table 6. Percentage of PT Sample Observations that are Outside of T $\pm 2\sigma_{pooled}$ and $\pm 2\sigma_{pooled}$**

| Analyte | Percentage of Outliers | % Outside T$\pm 2\sigma_{pooled}$ | | % Outside $\pm 2\sigma_{pooled}$ | |
|---|---|---|---|---|---|
| | | Including Outliers | Excluding Outliers | Including Outliers | Excluding Outliers |
| Arsenic | 2.2% | 5.8% | 3.6% | 8.0% | 5.8% |
| Nickel | 2.7% | 8.0% | 5.3% | 7.5% | 4.8% |
| Benzene | 1.0% | 4.3% | 3.3% | 4.0% | 3.0% |
| 1,3-butadiene | 2.8% | 10.0% | 7.2% | 8.4% | 5.6% |
| Formaldehyde | 1.7% | 16.1% | 14.4% | 7.4% | 5.7% |
| Acrolein | 1.3% | 9.2% | 7.9% | 7.2% | 5.9% |

## 3.3 Phase II Conclusions

The scatterplots in Appendix C demonstrate that the batch-specific laboratory averages are a good measure of the central tendency of the PT sample measurements, as the black line tends to go through the sample measurements close to their center over the range of laboratory averages. Thus, it is a good representation of the target value for a batch of PT samples. Frequently, a similar line representing the measure obtained from the single PT provider laboratory did not resemble a line of symmetry for the laboratory measurements, and thus, a greater percentage of the measurements fell outside of the current MQO acceptable range of ±25% of T.

For the six analytes considered in Phase II, basing the alternative MQO on $2\sigma_{pooled}$ (and thus making it analyte-specific) typically led to a larger range of acceptability compared to when a constant criterion of 25% was used. While the larger range results in a lower percentage of sample measurements failing to achieve the MQO, the range is not so large that it includes nearly every measurement for each analyte. When the lab average was used as the target value, from 3% to 6% of the measurements failed to achieve the analyte-specific MQO even when outliers were not considered, meaning that poor performance at laboratories, if present, should still be discovered during PT testing. Up to 14% of the measurements failed to achieve the analyte-specific MQO when the target T was used instead.

Overall, the results of Phase II suggest that an analyte-specific MQO that utilizes the lab average as the target sample concentration (i.e., is based on $PD_2$) is feasible to consider and may be more appropriate than the current MQO.

## 4.0 Overall Conclusions

This analysis provides useful information for investigating the extent to which the historic NATTS MQO for laboratory bias, ± 25% of a stated "true" value of the PT samples, should be revised to account for a different approach to determining the target concentration in each batch, and/or to make it analyte-specific by considering the actual ability of the participating laboratories to implement the analytical methods over time. The alternative MQO considered in this analysis was two times the pooled standard deviation of the percent difference from the lab average, after removing batch-to-batch variability. The outcome of this analysis suggests the following:

- The MQOs should be analyte-specific, or specific to an analyte class, rather than being a fixed percentage of a stated "true" value.
- Assessments of bias through analysis of PT samples should account for the expected range of uncertainty associated with the given analyte, which is itself a function of the capabilities of the different analytical methods for VOCs, metals, and carbonyls.
- Laboratory performance can vary over time, and thus, the uncertainty estimates used in the MQO should be reassessed and modified periodically.
- Setting the target value to equal the laboratory average (perhaps weighted by the number of samples contributed by the laboratory) may be a more appropriate measure than the current approach of assigning the target using results from the PT provider lab.

# APPENDIX A

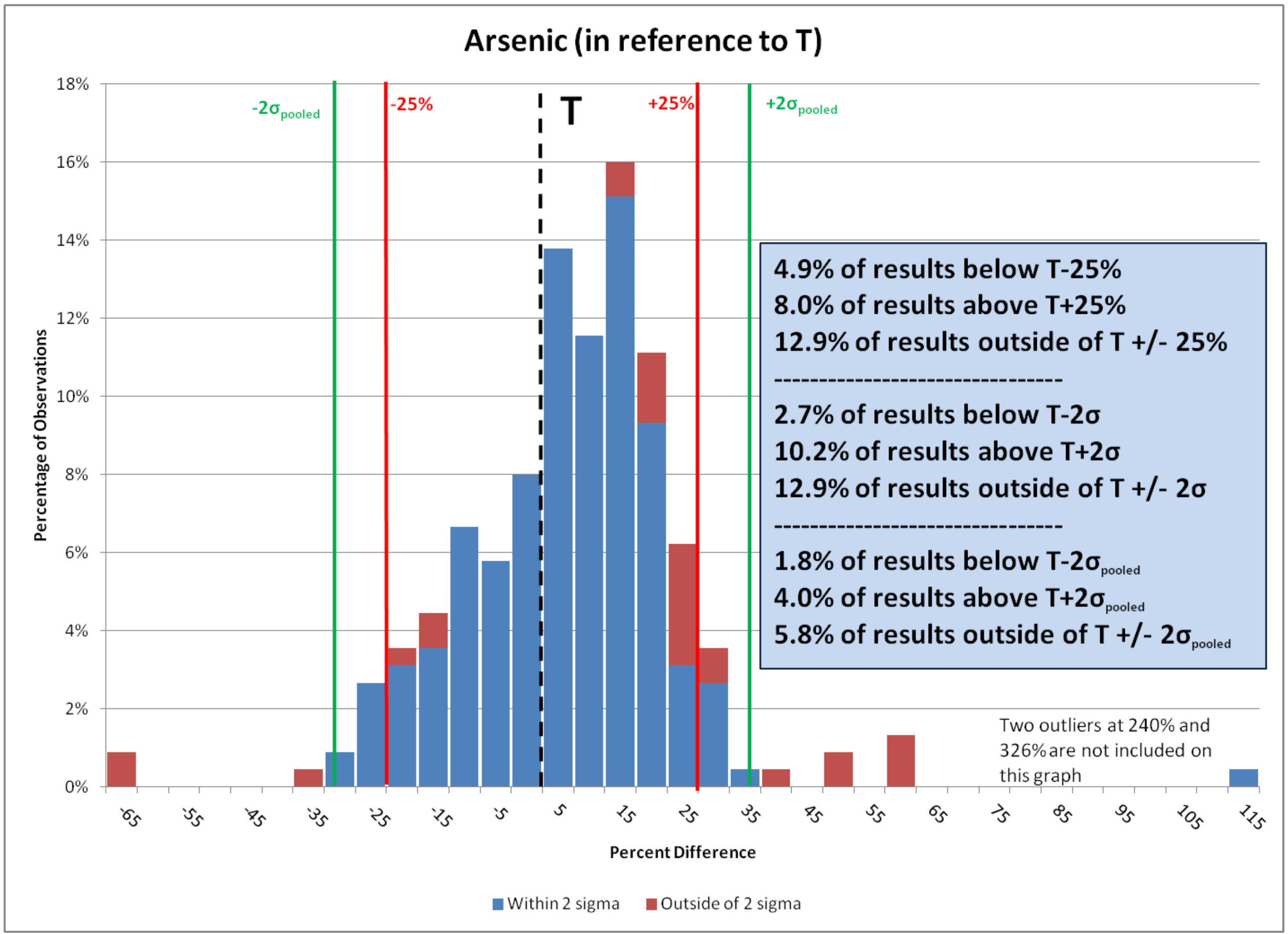
## BAR CHARTS OF PERCENT DIFFERENCE FROM TARGET, FOR 23 ANALYTES INCLUDED IN PROFICIENCY TEST SAMPLES FOR THE NATTS PROGRAM

Top chart on each page:

- o Target = T (determined from a single lab or spike level)
- o Percent difference = $PD_1$

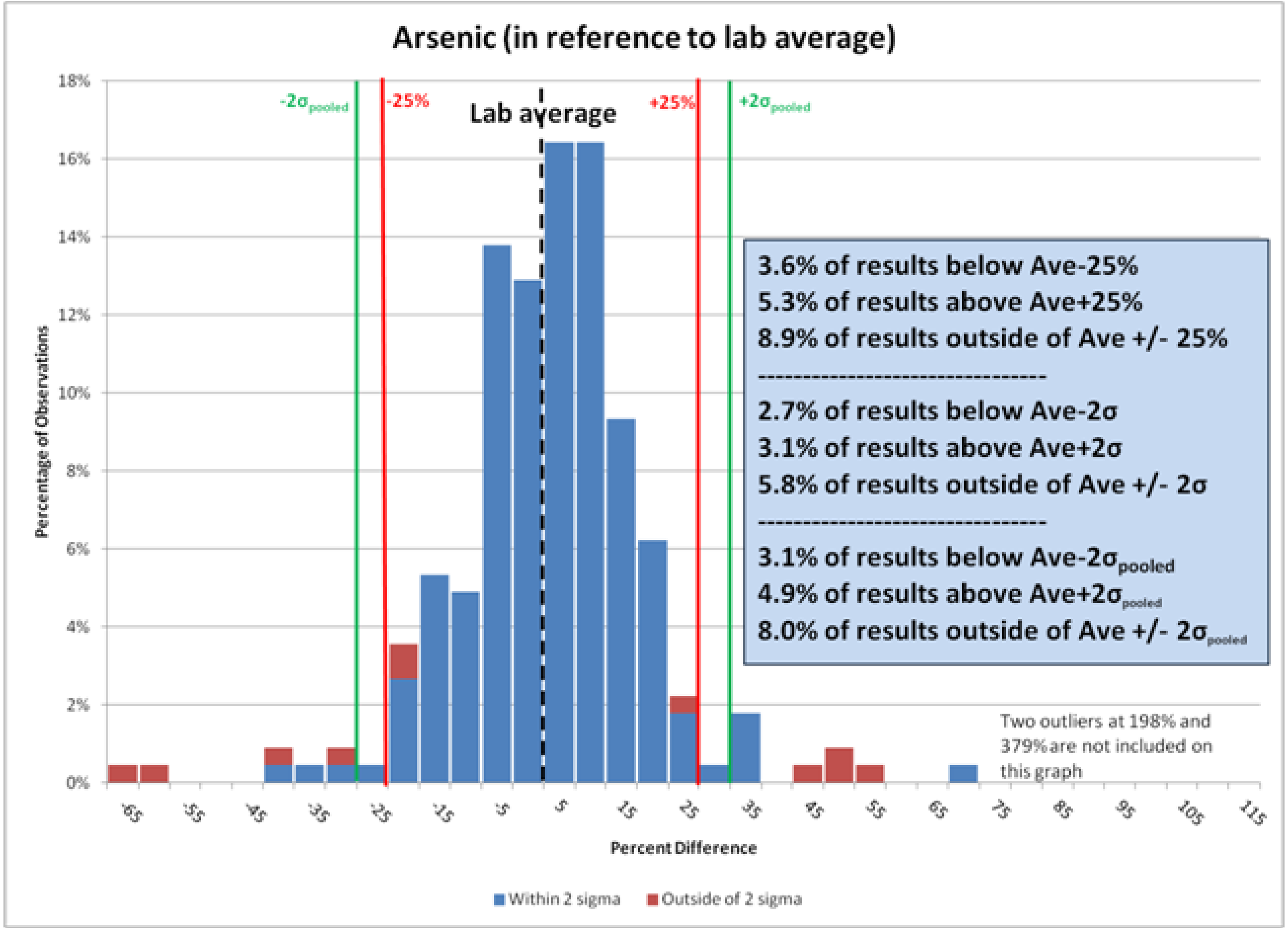
Bottom chart on each page:

- o Target = batch-specific average across labs
- o Percent difference = $PD_2$



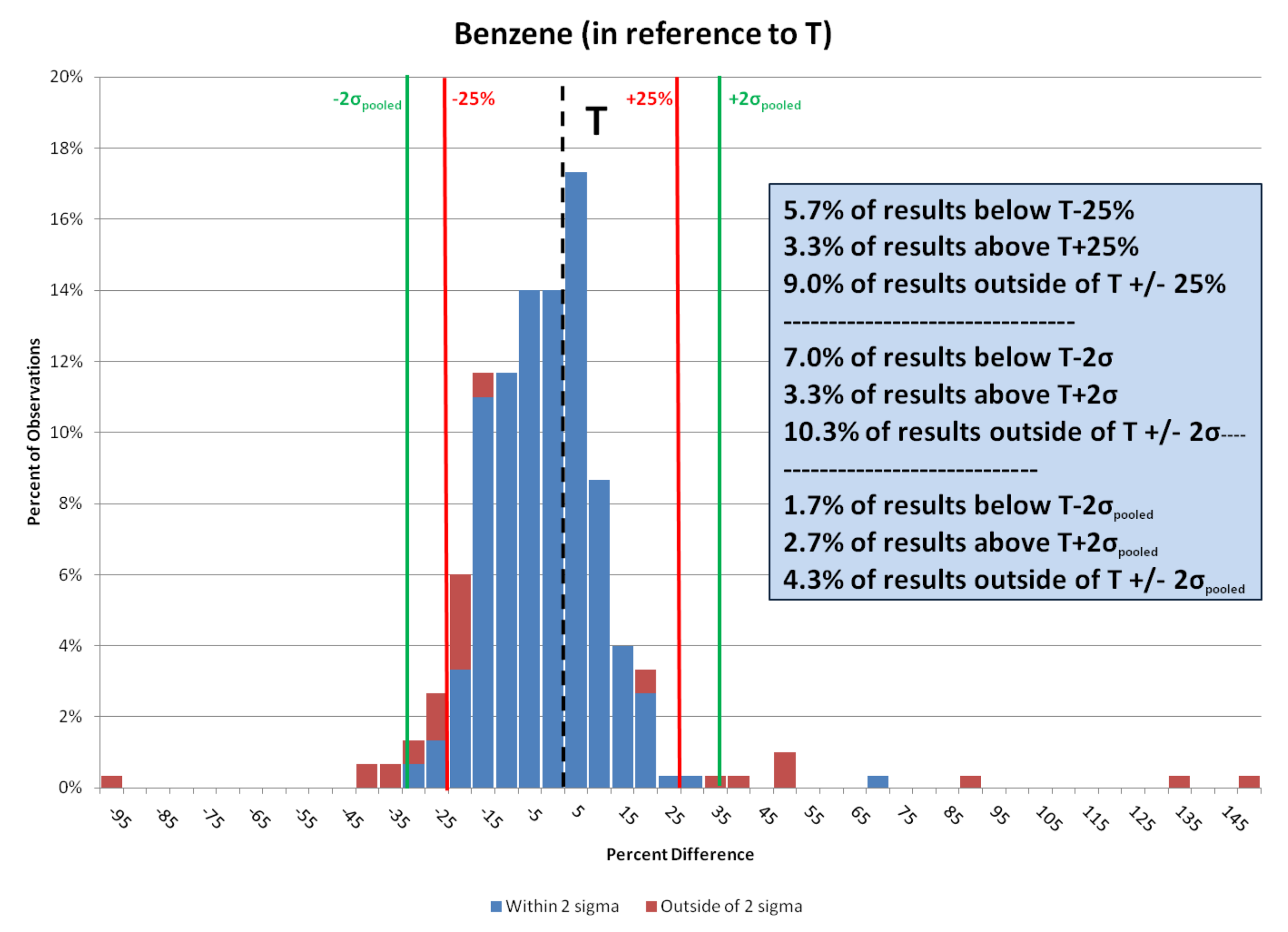
4.9% of results below T-25%
8.0% of results above T+25%
12.9% of results outside of T +/- 25%
--------------------------------
2.7% of results below T-2σ
10.2% of results above T+2σ
12.9% of results outside of T +/- 2σ
--------------------------------
1.8% of results below $T-2\sigma_{pooled}$
4.0% of results above $T+2\sigma_{pooled}$
5.8% of results outside of T +/- $2\sigma_{pooled}$

Two outliers at 240% and 326% are not included on this graph



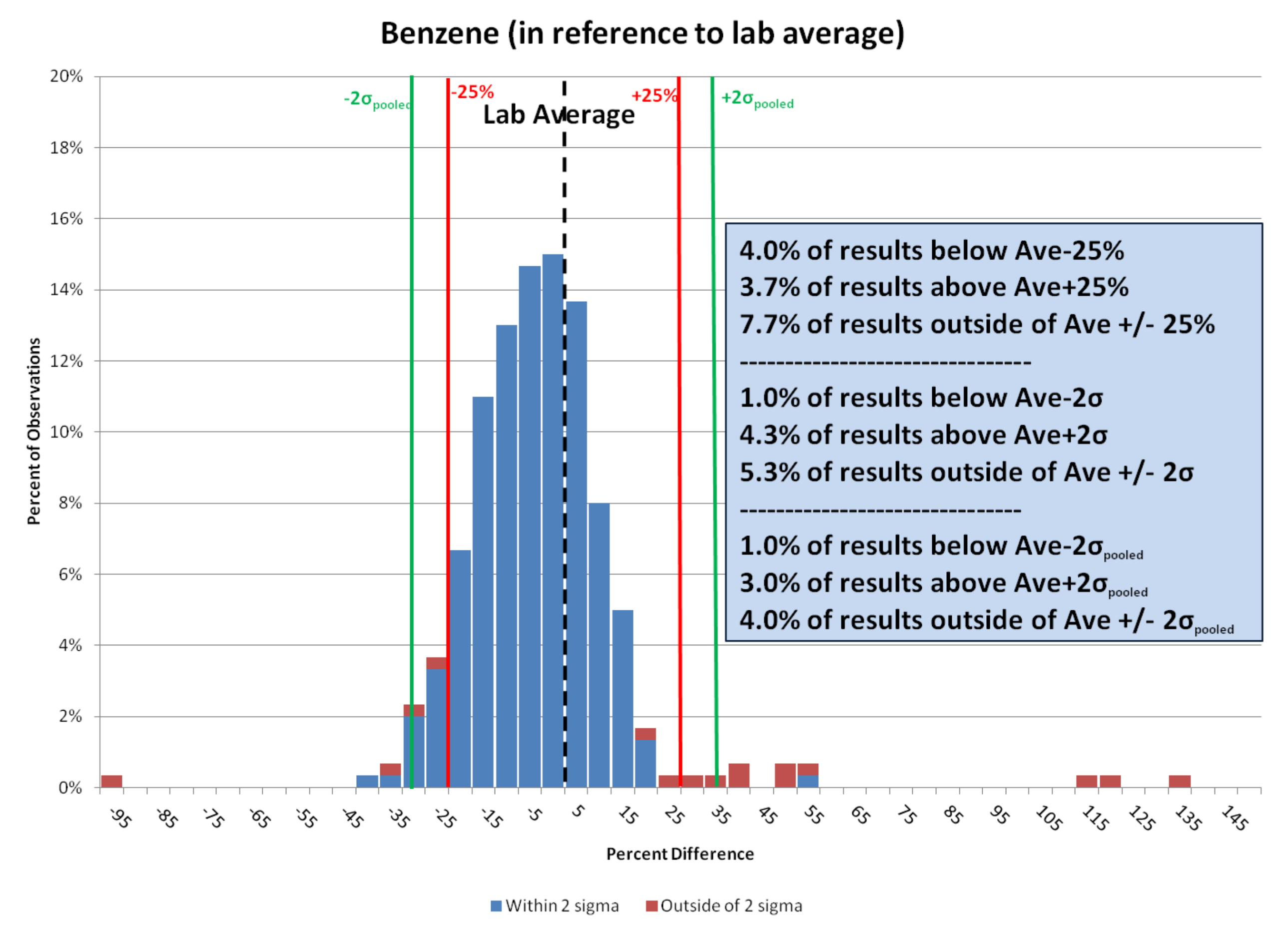
3.6% of results below Ave-25%
5.3% of results above Ave+25%
8.9% of results outside of Ave +/- 25%
--------------------------------
2.7% of results below Ave-2σ
3.1% of results above Ave+2σ
5.8% of results outside of Ave +/- 2σ
--------------------------------
3.1% of results below $Ave-2\sigma_{pooled}$
4.9% of results above $Ave+2\sigma_{pooled}$
8.0% of results outside of Ave +/- $2\sigma_{pooled}$

Two outliers at 198% and 379% are not included on this graph



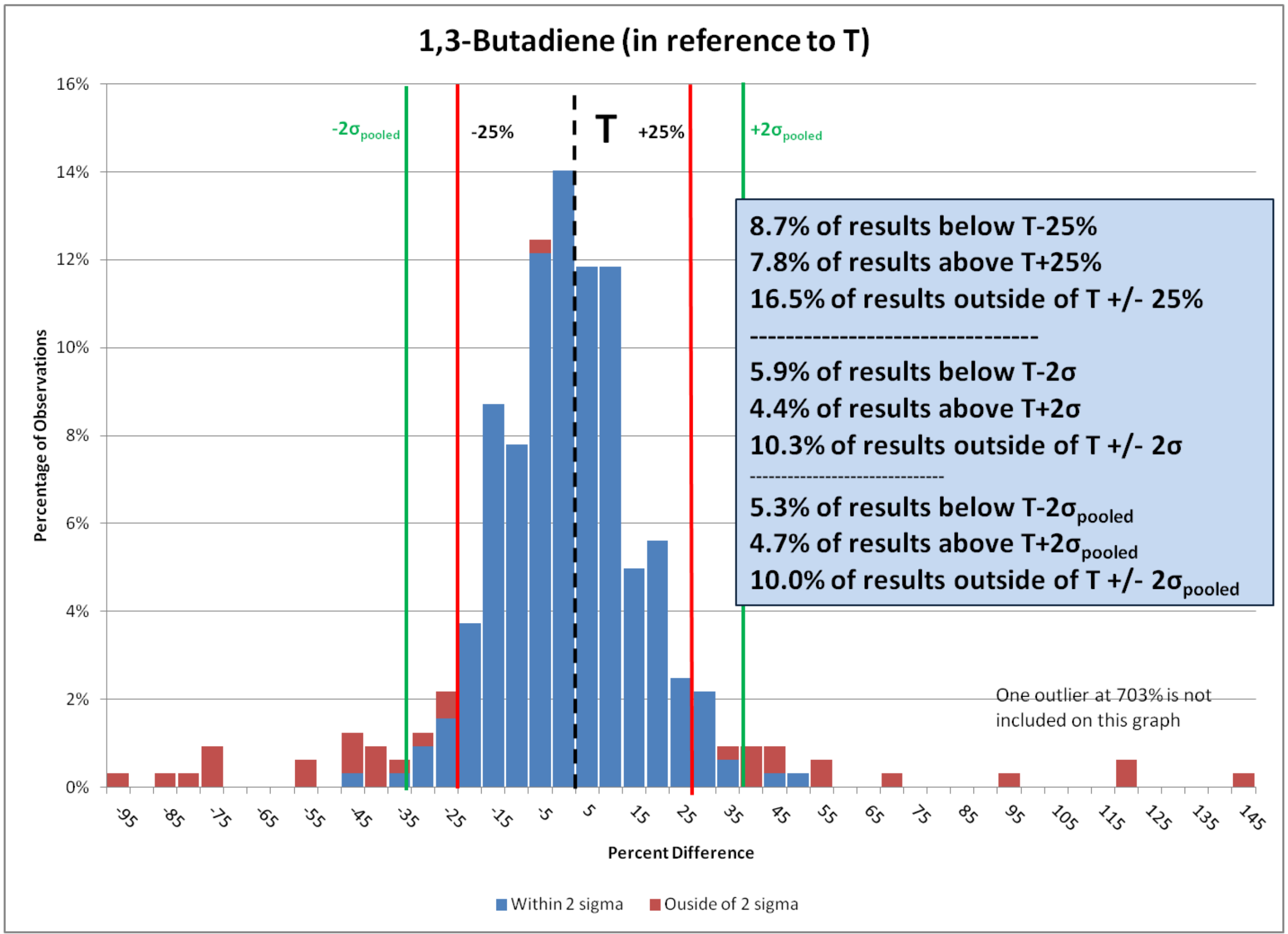
Benzene (in reference to T)

5.7% of results below T-25%
3.3% of results above T+25%
9.0% of results outside of T +/- 25%
--------------------------------
7.0% of results below T-2σ
3.3% of results above T+2σ
10.3% of results outside of T +/- 2σ----
----------------------------
1.7% of results below T-$2\sigma_{pooled}$
2.7% of results above T+$2\sigma_{pooled}$
4.3% of results outside of T +/- $2\sigma_{pooled}$



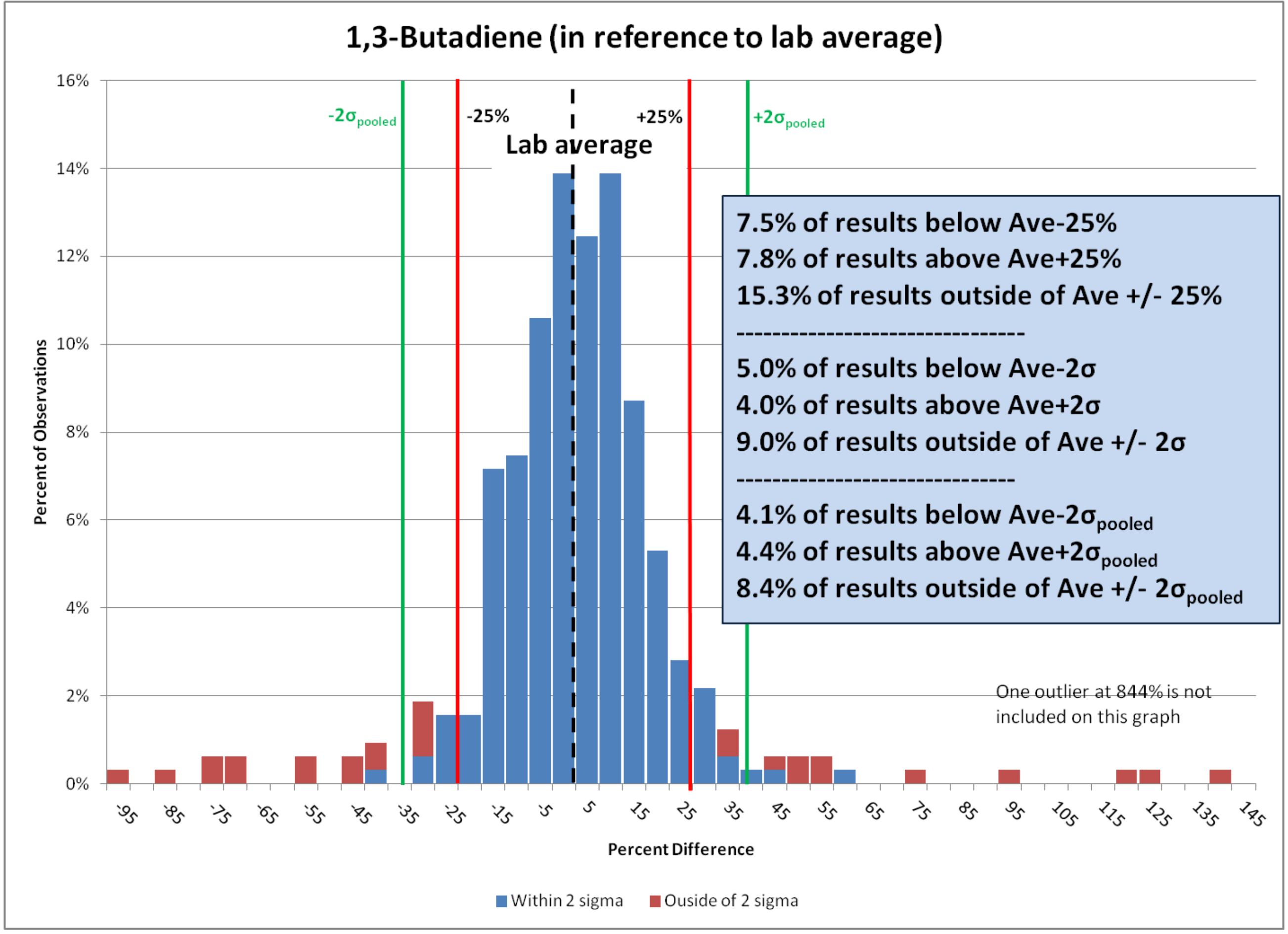
Benzene (in reference to lab average)

4.0% of results below Ave-25%
3.7% of results above Ave+25%
7.7% of results outside of Ave +/- 25%
--------------------------------
1.0% of results below Ave-2σ
4.3% of results above Ave+2σ
5.3% of results outside of Ave +/- 2σ
-------------------------------
1.0% of results below Ave-$2\sigma_{pooled}$
3.0% of results above Ave+$2\sigma_{pooled}$
4.0% of results outside of Ave +/- $2\sigma_{pooled}$


1,3-Butadiene (in reference to T)

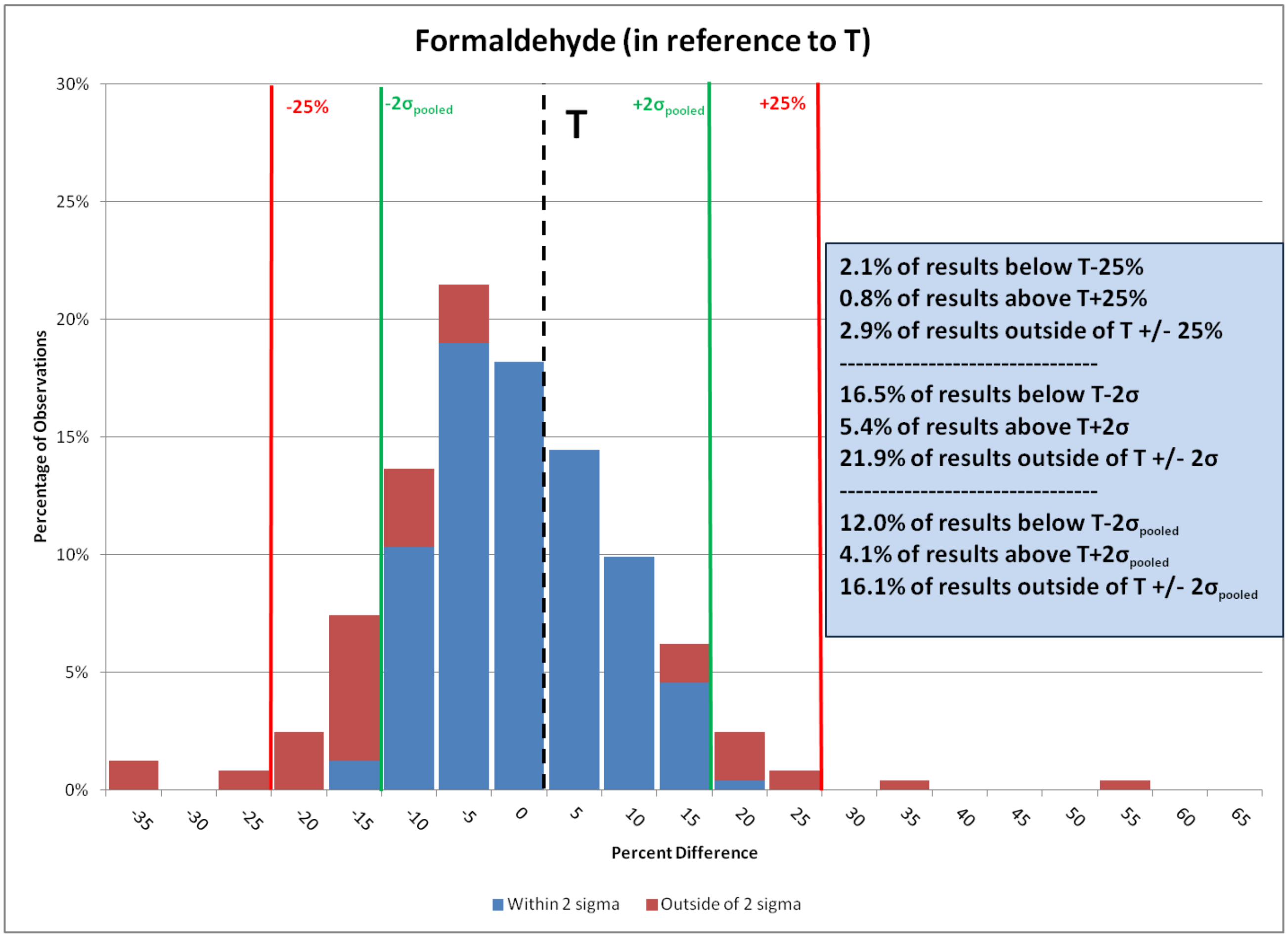
8.7% of results below T-25%
7.8% of results above T+25%
16.5% of results outside of T +/- 25%
--------------------------------
5.9% of results below T-2σ
4.4% of results above T+2σ
10.3% of results outside of T +/- 2σ
------------------------------
5.3% of results below T-2$\sigma_{pooled}$
4.7% of results above T+2$\sigma_{pooled}$
10.0% of results outside of T +/- 2$\sigma_{pooled}$

One outlier at 703% is not included on this graph


1,3-Butadiene (in reference to lab average)

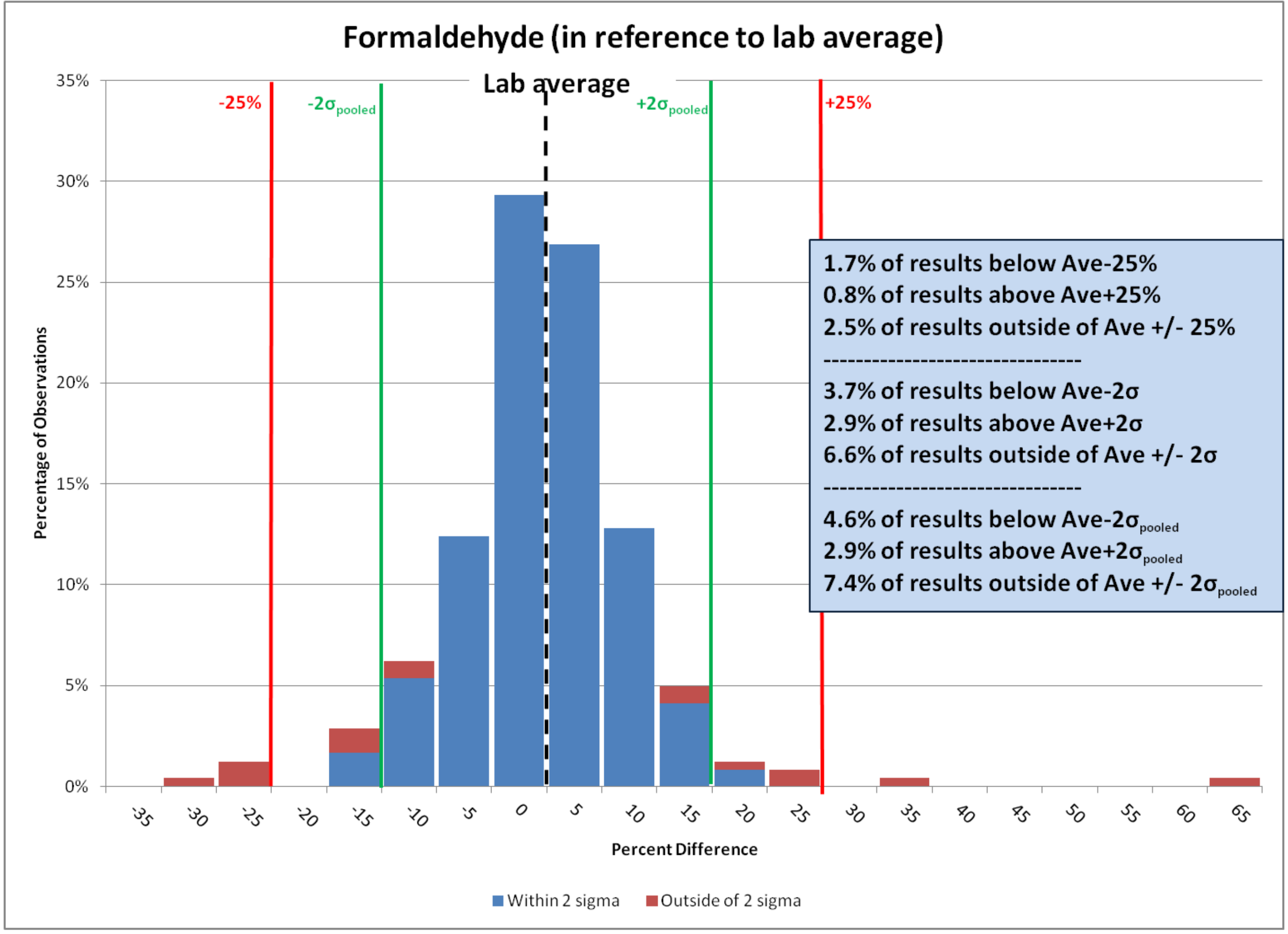
7.5% of results below Ave-25%
7.8% of results above Ave+25%
15.3% of results outside of Ave +/- 25%
--------------------------------
5.0% of results below Ave-2σ
4.0% of results above Ave+2σ
9.0% of results outside of Ave +/- 2σ
------------------------------
4.1% of results below Ave-2$\sigma_{pooled}$
4.4% of results above Ave+2$\sigma_{pooled}$
8.4% of results outside of Ave +/- 2$\sigma_{pooled}$

One outlier at 844% is not included on this graph



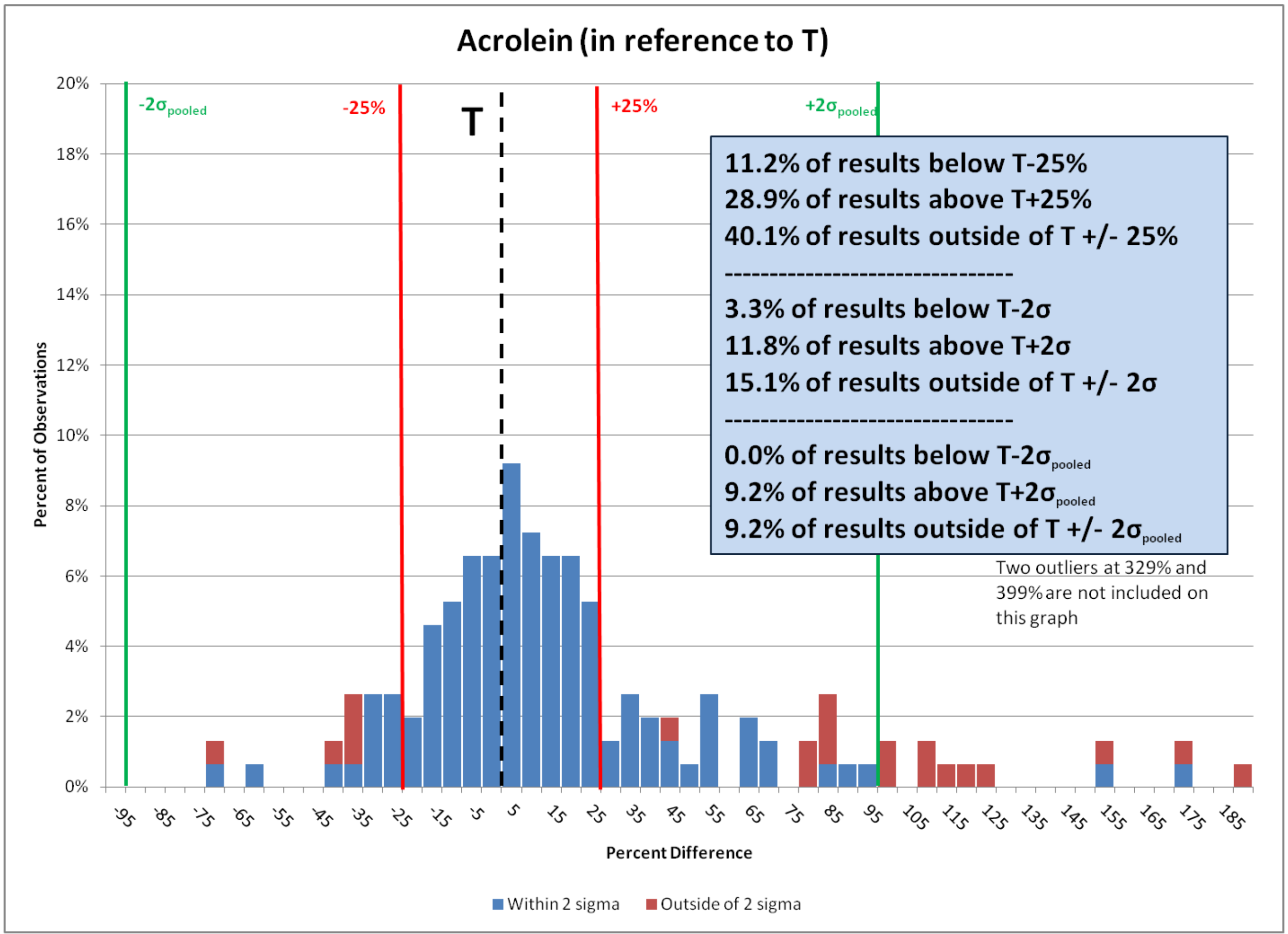
Formaldehyde (in reference to T)

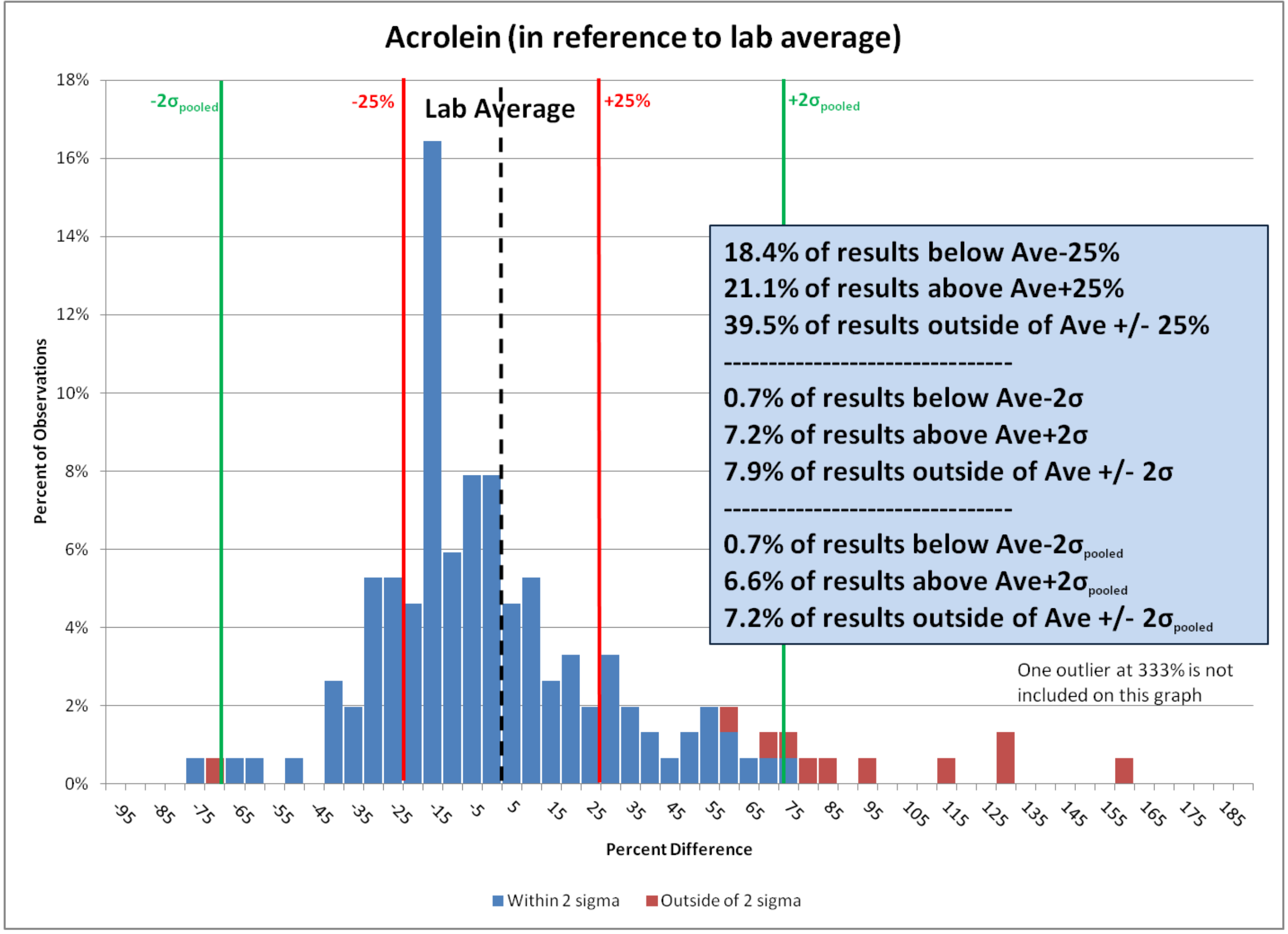
2.1% of results below T-25%
0.8% of results above T+25%
2.9% of results outside of T +/- 25%
--------------------------------
16.5% of results below T-2σ
5.4% of results above T+2σ
21.9% of results outside of T +/- 2σ
--------------------------------
12.0% of results below T-$2\sigma_{pooled}$
4.1% of results above T+$2\sigma_{pooled}$
16.1% of results outside of T +/- $2\sigma_{pooled}$

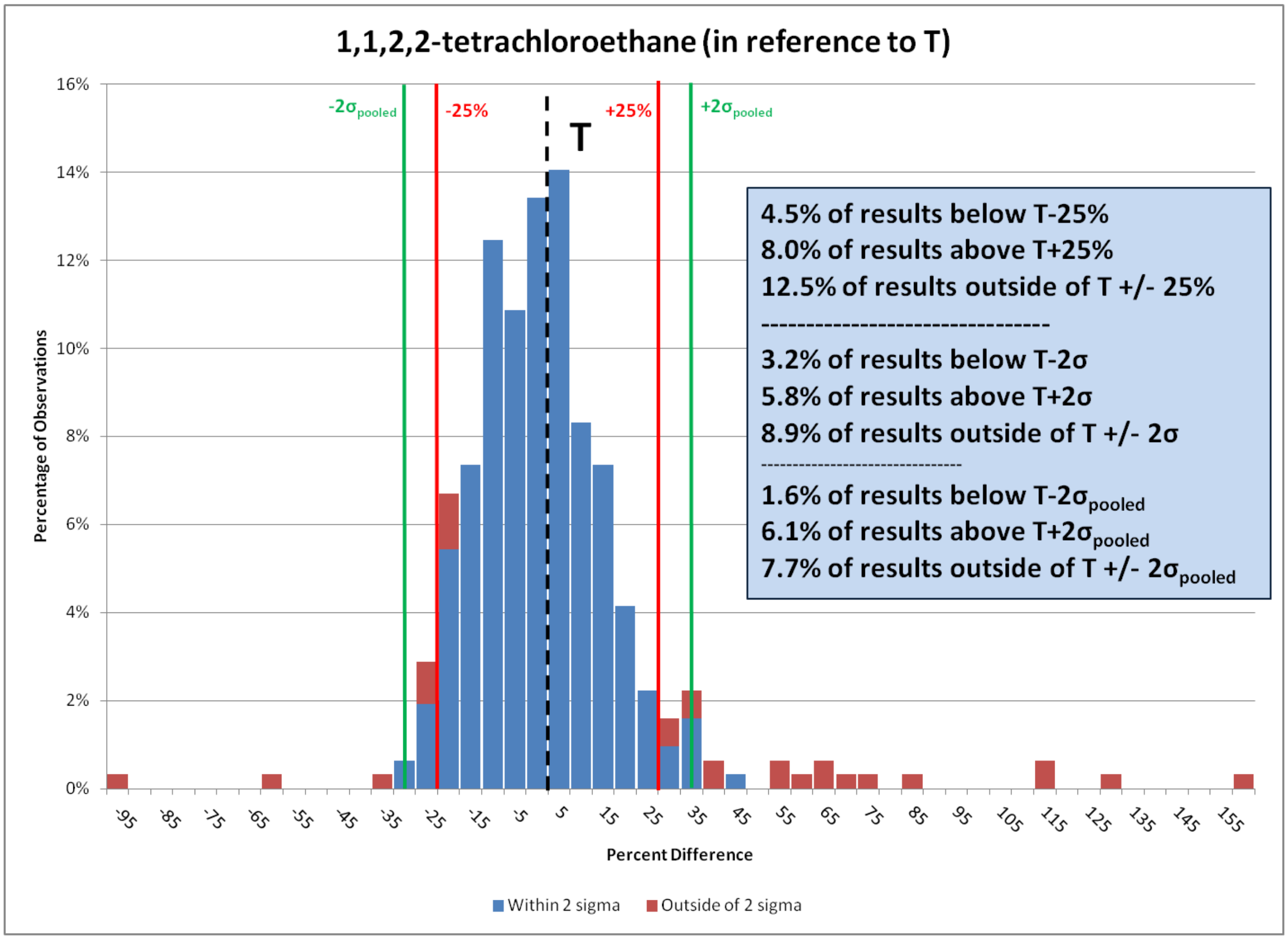


Formaldehyde (in reference to lab average)

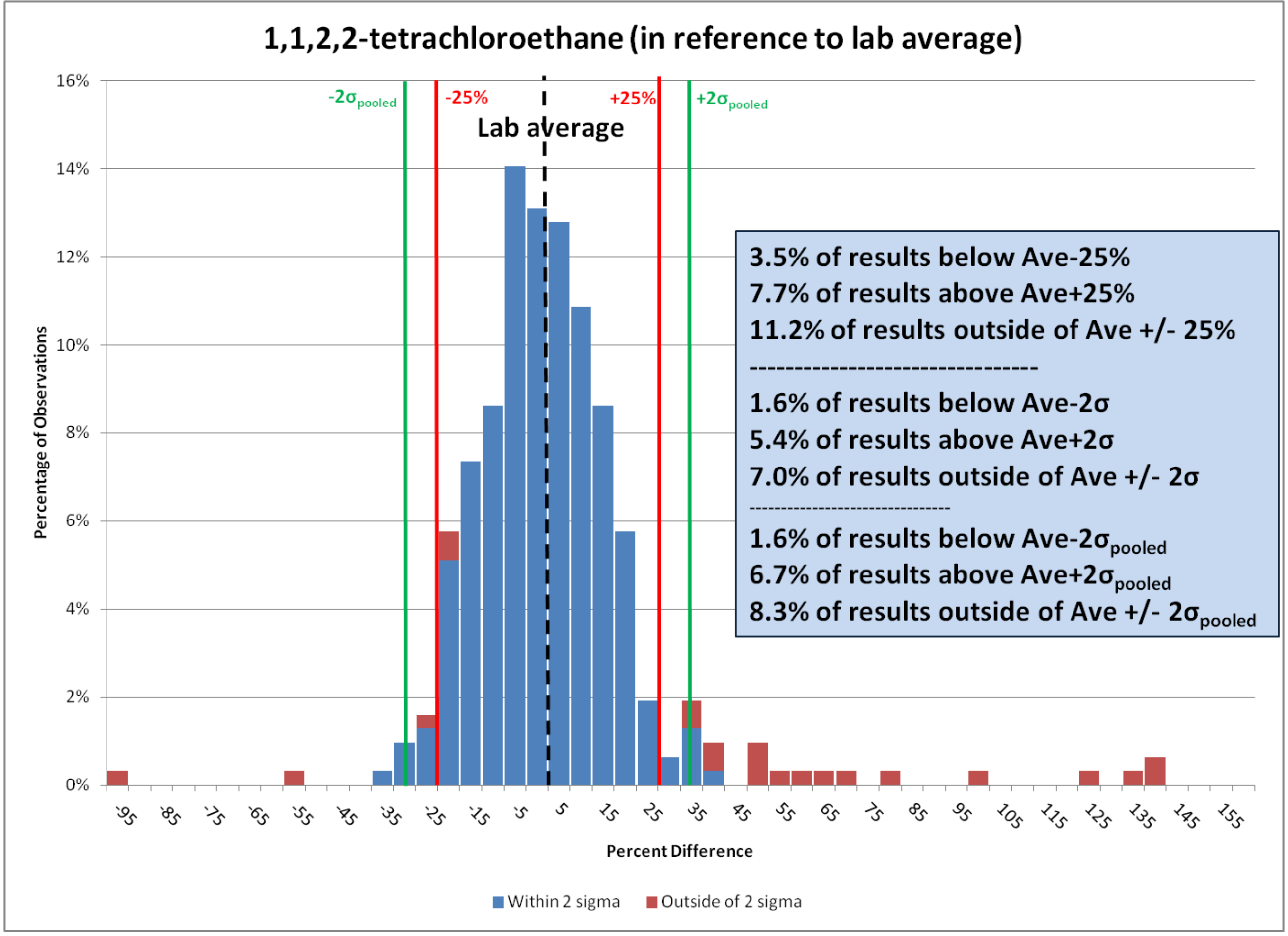
1.7% of results below Ave-25%
0.8% of results above Ave+25%
2.5% of results outside of Ave +/- 25%
--------------------------------
3.7% of results below Ave-2σ
2.9% of results above Ave+2σ
6.6% of results outside of Ave +/- 2σ
--------------------------------
4.6% of results below Ave-$2\sigma_{pooled}$
2.9% of results above Ave+$2\sigma_{pooled}$
7.4% of results outside of Ave +/- $2\sigma_{pooled}$


Acrolein (in reference to T)
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
Percent of Observations
$-2\sigma_{pooled}$
-25%
T
+25%
$+2\sigma_{pooled}$
11.2% of results below T-25%
28.9% of results above T+25%
40.1% of results outside of T +/- 25%
--------------------------------
3.3% of results below T-2σ
11.8% of results above T+2σ
15.1% of results outside of T +/- 2σ
--------------------------------
0.0% of results below T-$2\sigma_{pooled}$
9.2% of results above T+$2\sigma_{pooled}$
9.2% of results outside of T +/- $2\sigma_{pooled}$
Two outliers at 329% and 399% are not included on this graph
-95 -85 -75 -65 -55 -45 -35 -25 -15 -5 5 15 25 35 45 55 65 75 85 95 105 115 125 135 145 155 165 175 185
Percent Difference
Within 2 sigma
Outside of 2 sigma

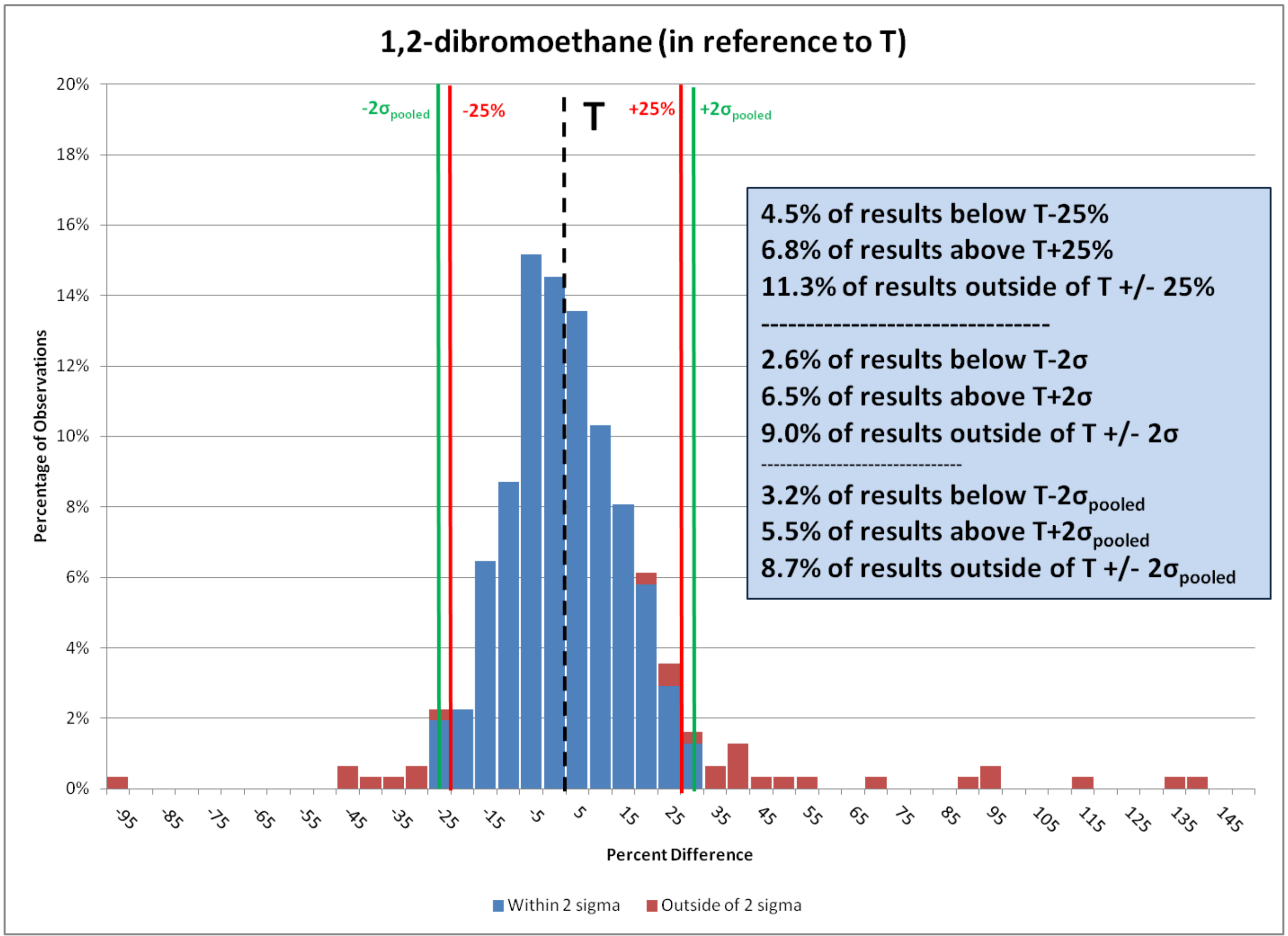

Acrolein (in reference to lab average)
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
Percent of Observations
$-2\sigma_{pooled}$
-25%
Lab Average
+25%
$+2\sigma_{pooled}$
18.4% of results below Ave-25%
21.1% of results above Ave+25%
39.5% of results outside of Ave +/- 25%
--------------------------------
0.7% of results below Ave-2σ
7.2% of results above Ave+2σ
7.9% of results outside of Ave +/- 2σ
--------------------------------
0.7% of results below Ave-$2\sigma_{pooled}$
6.6% of results above Ave+$2\sigma_{pooled}$
7.2% of results outside of Ave +/- $2\sigma_{pooled}$
One outlier at 333% is not included on this graph
-95 -85 -75 -65 -55 -45 -35 -25 -15 -5 5 15 25 35 45 55 65 75 85 95 105 115 125 135 145 155 165 175 185
Percent Difference
Within 2 sigma
Outside of 2 sigma

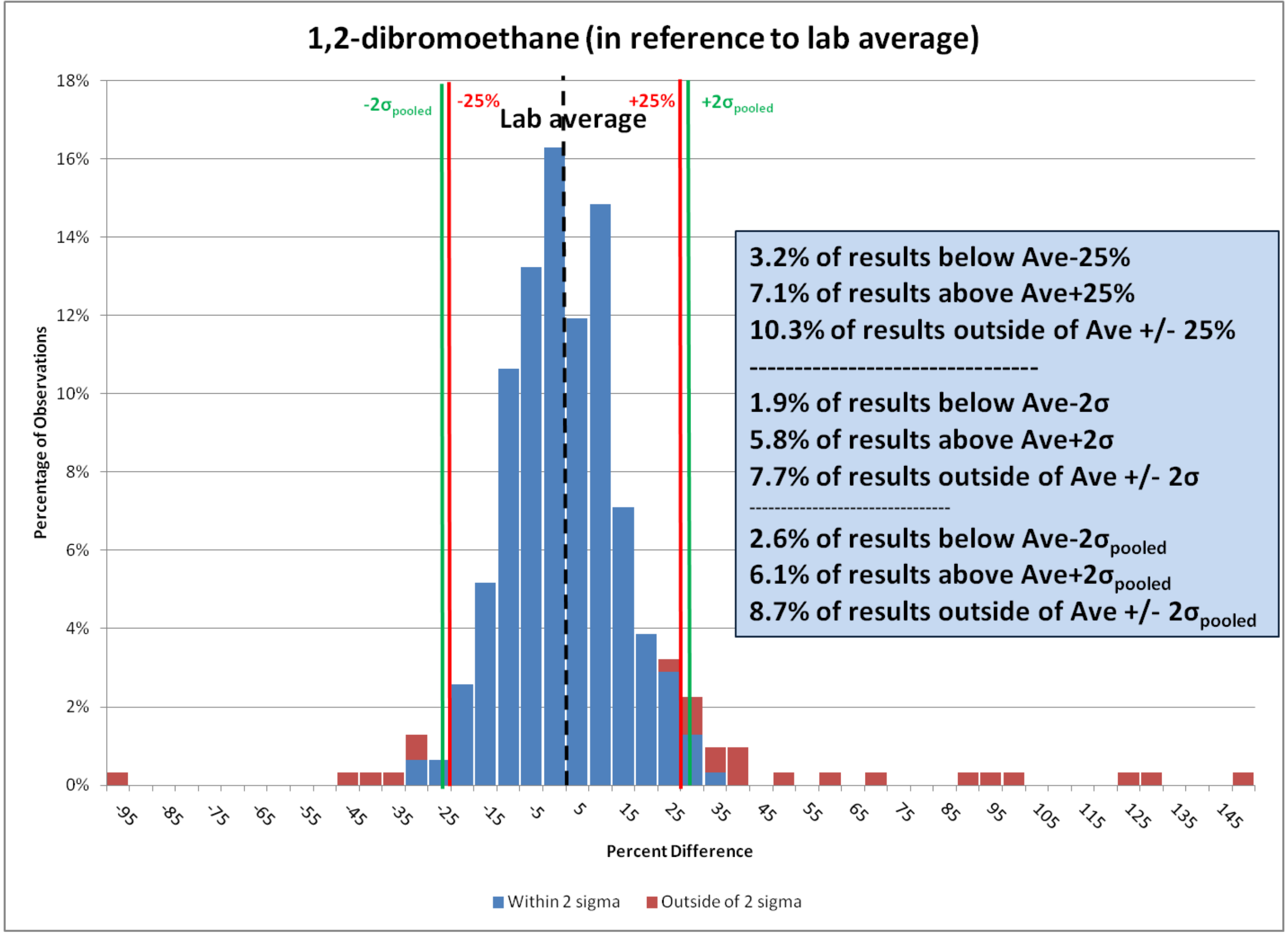


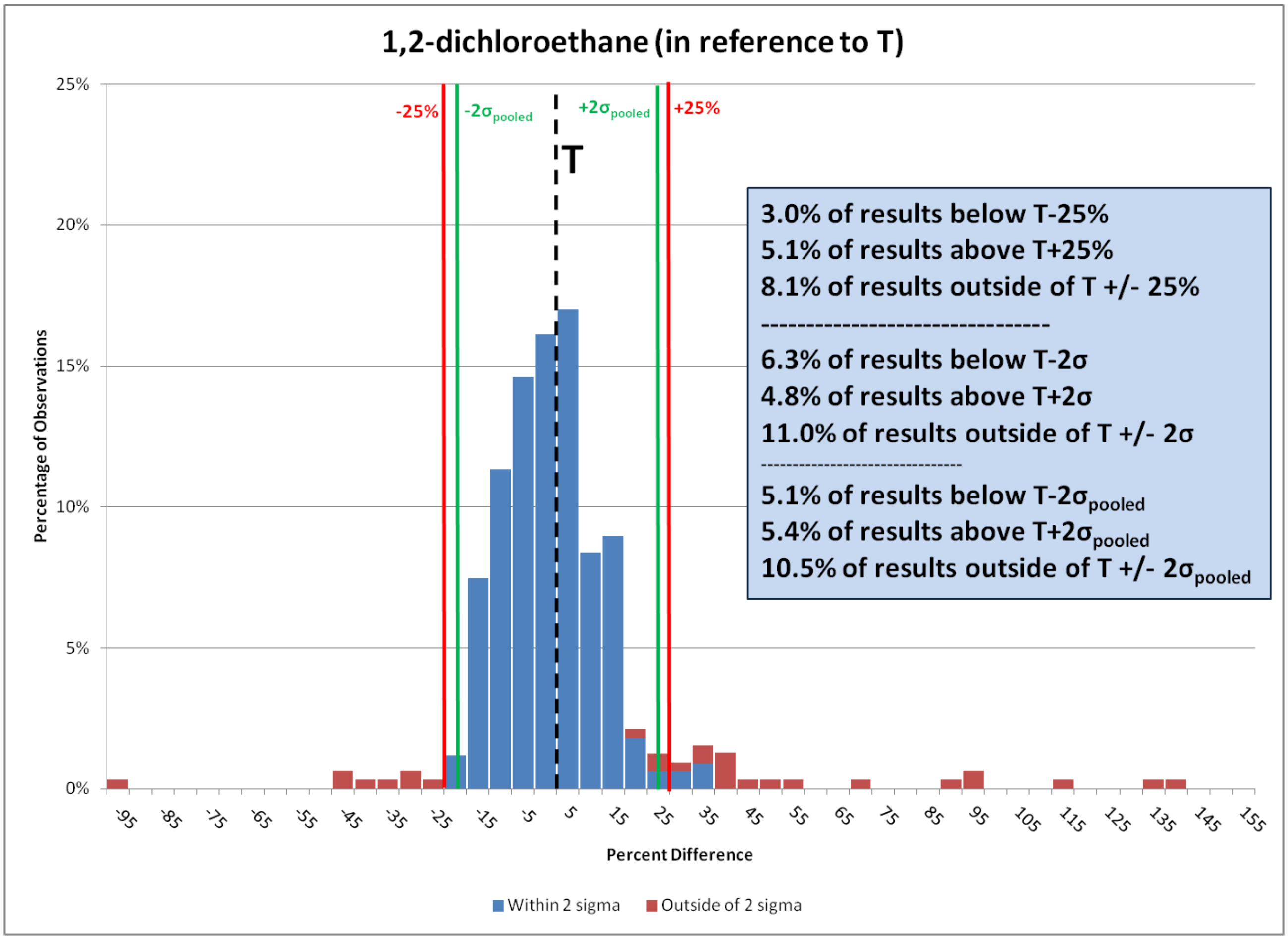
4.5% of results below T-25%
8.0% of results above T+25%
12.5% of results outside of T +/- 25%
--------------------------------
3.2% of results below T-2σ
5.8% of results above T+2σ
8.9% of results outside of T +/- 2σ
--------------------------------
1.6% of results below T-$2\sigma_{pooled}$
6.1% of results above T+$2\sigma_{pooled}$
7.7% of results outside of T +/- $2\sigma_{pooled}$



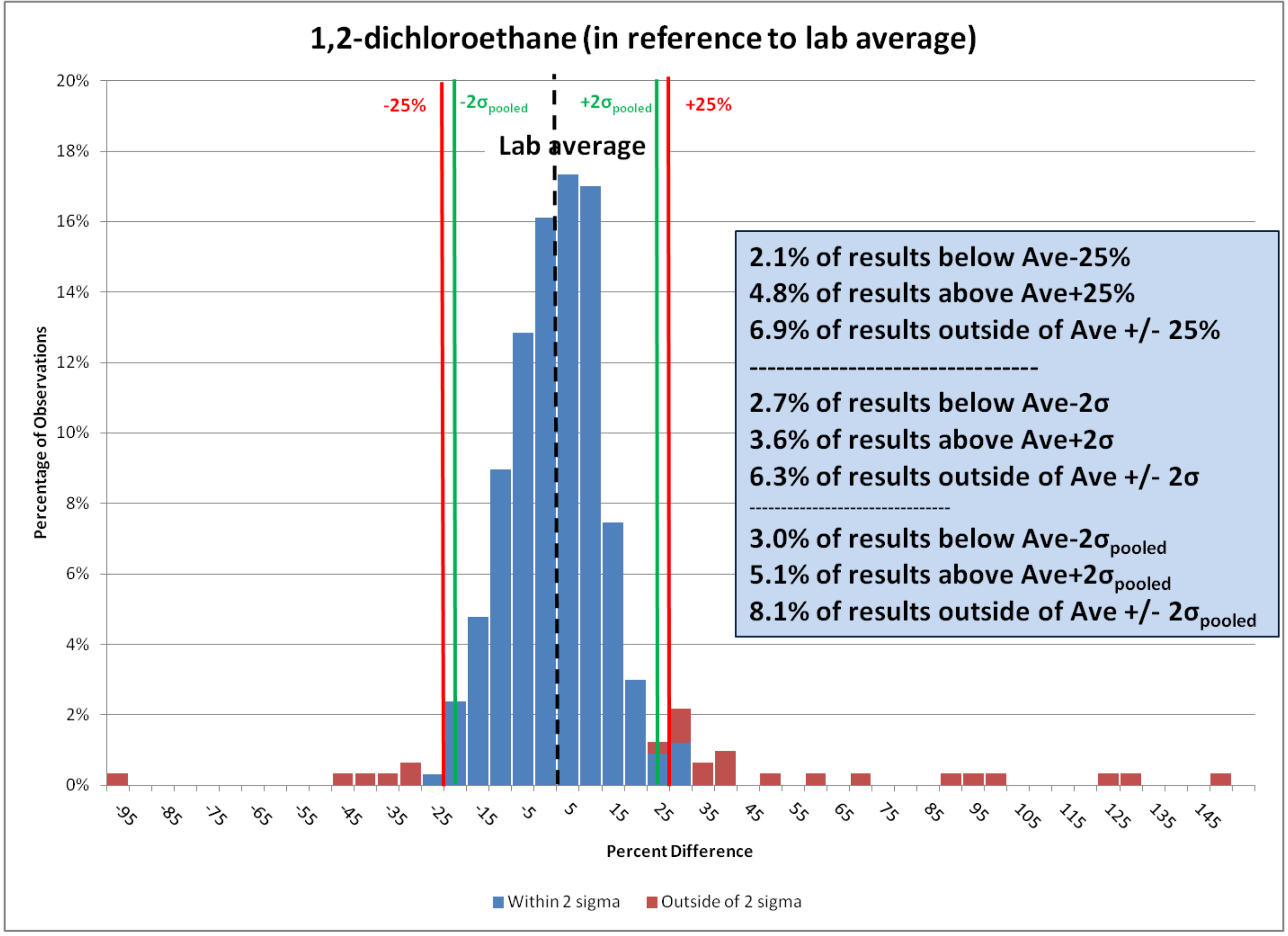
3.5% of results below Ave-25%
7.7% of results above Ave+25%
11.2% of results outside of Ave +/- 25%
--------------------------------
1.6% of results below Ave-2σ
5.4% of results above Ave+2σ
7.0% of results outside of Ave +/- 2σ
--------------------------------
1.6% of results below Ave-$2\sigma_{pooled}$
6.7% of results above Ave+$2\sigma_{pooled}$
8.3% of results outside of Ave +/- $2\sigma_{pooled}$


1,2-dibromoethane (in reference to T)

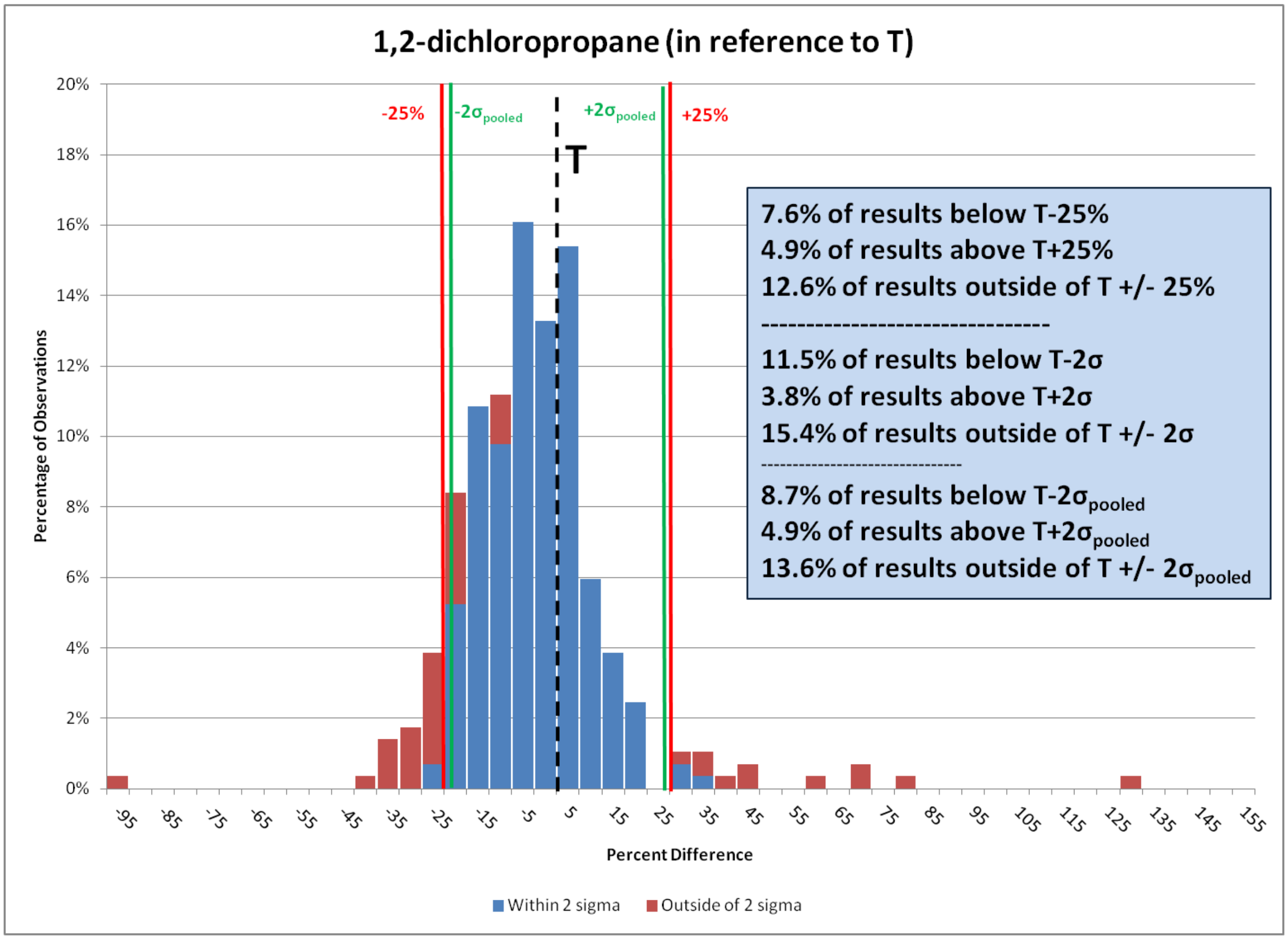

1,2-dibromoethane (in reference to lab average)



3.0% of results below T-25%
5.1% of results above T+25%
8.1% of results outside of T +/- 25%
--------------------------------
6.3% of results below T-2σ
4.8% of results above T+2σ
11.0% of results outside of T +/- 2σ
--------------------------------
5.1% of results below T-$2\sigma_{pooled}$
5.4% of results above T+$2\sigma_{pooled}$
10.5% of results outside of T +/- $2\sigma_{pooled}$

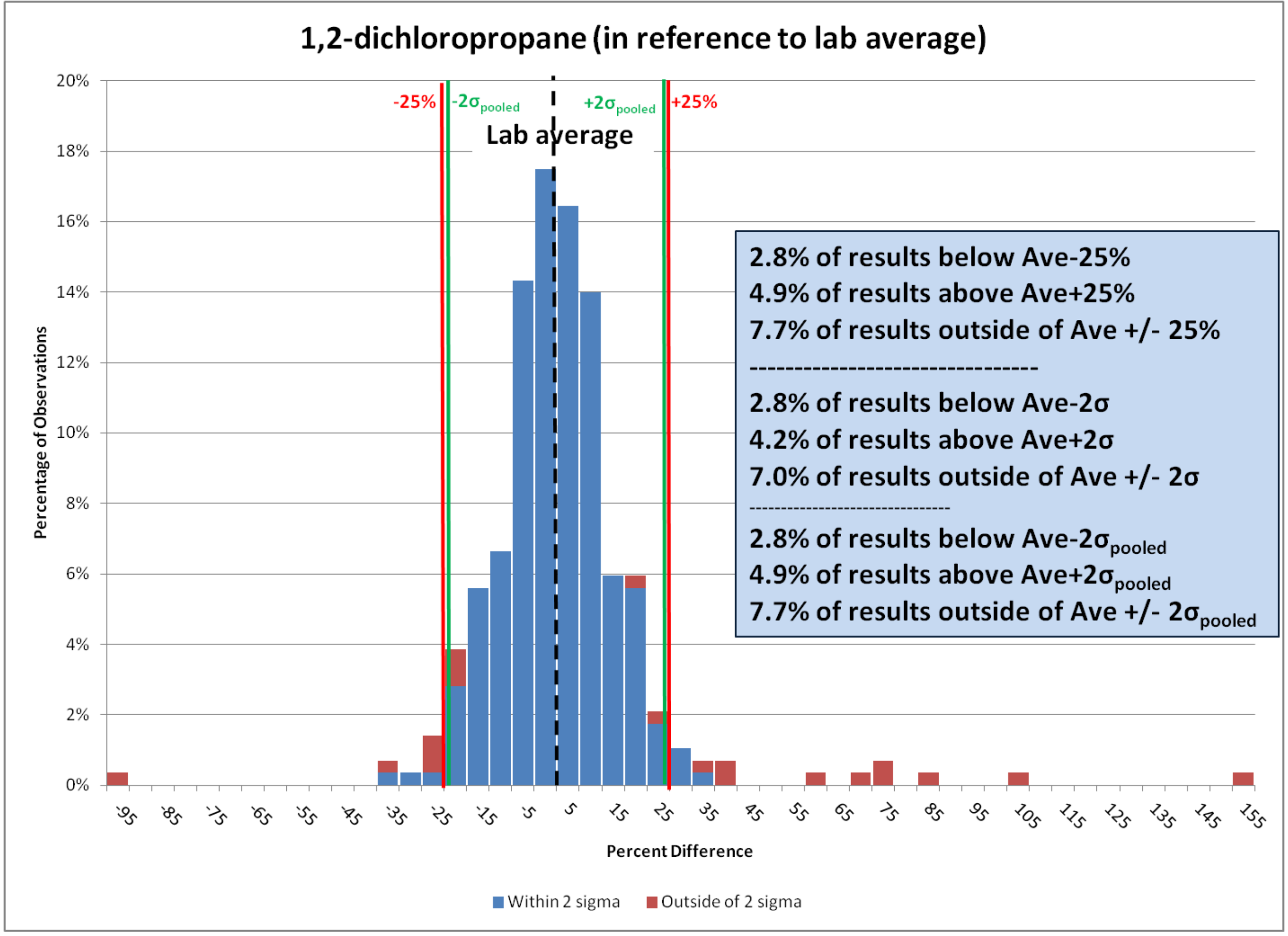


2.1% of results below Ave-25%
4.8% of results above Ave+25%
6.9% of results outside of Ave +/- 25%
--------------------------------
2.7% of results below Ave-2σ
3.6% of results above Ave+2σ
6.3% of results outside of Ave +/- 2σ
--------------------------------
3.0% of results below Ave-$2\sigma_{pooled}$
5.1% of results above Ave+$2\sigma_{pooled}$
8.1% of results outside of Ave +/- $2\sigma_{pooled}$



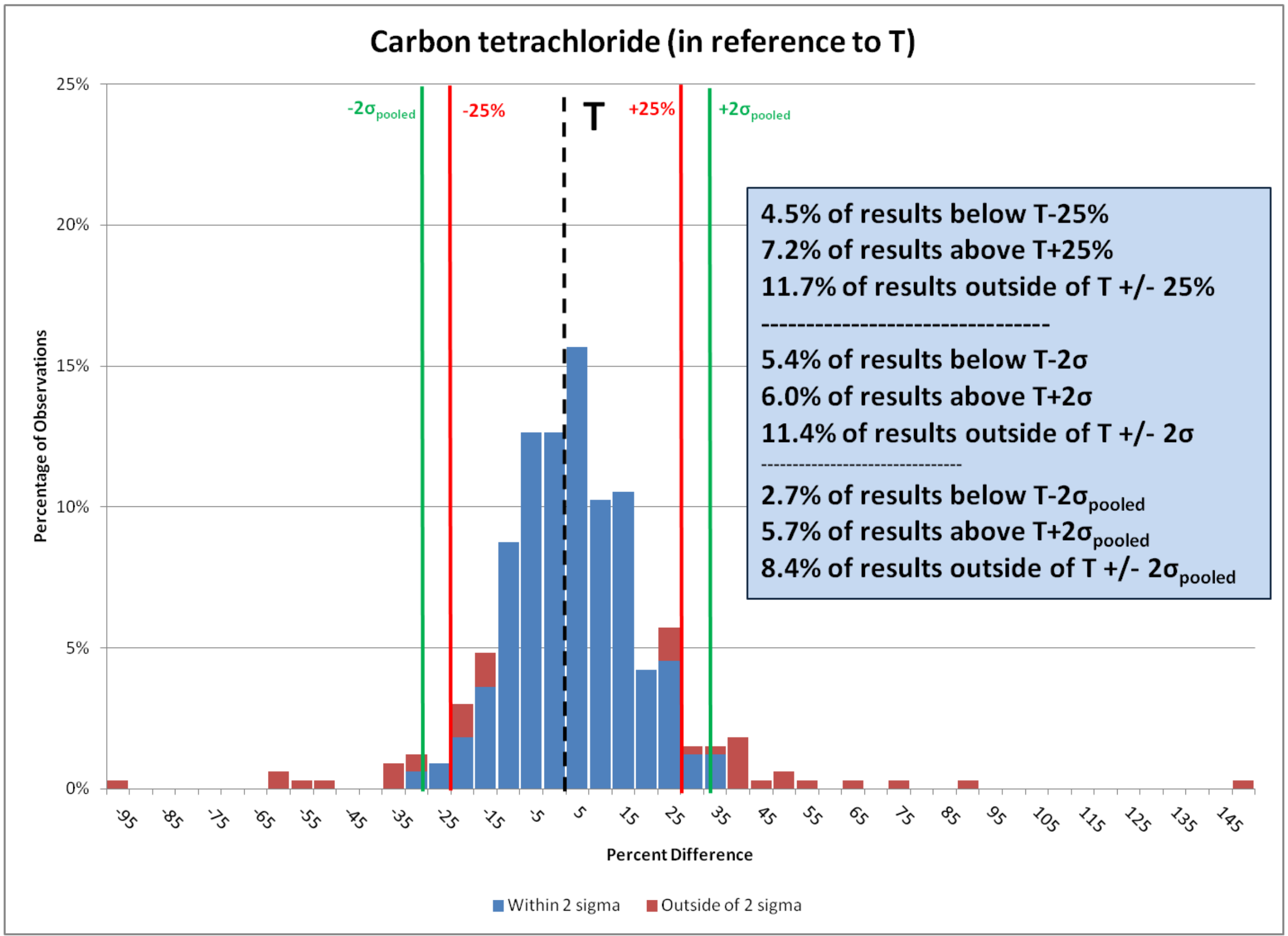
1,2-dichloropropane (in reference to T)

7.6% of results below T-25%
4.9% of results above T+25%
12.6% of results outside of T +/- 25%
--------------------------------
11.5% of results below T-2σ
3.8% of results above T+2σ
15.4% of results outside of T +/- 2σ
--------------------------------
8.7% of results below T-2$\sigma_{pooled}$
4.9% of results above T+2$\sigma_{pooled}$
13.6% of results outside of T +/- 2$\sigma_{pooled}$



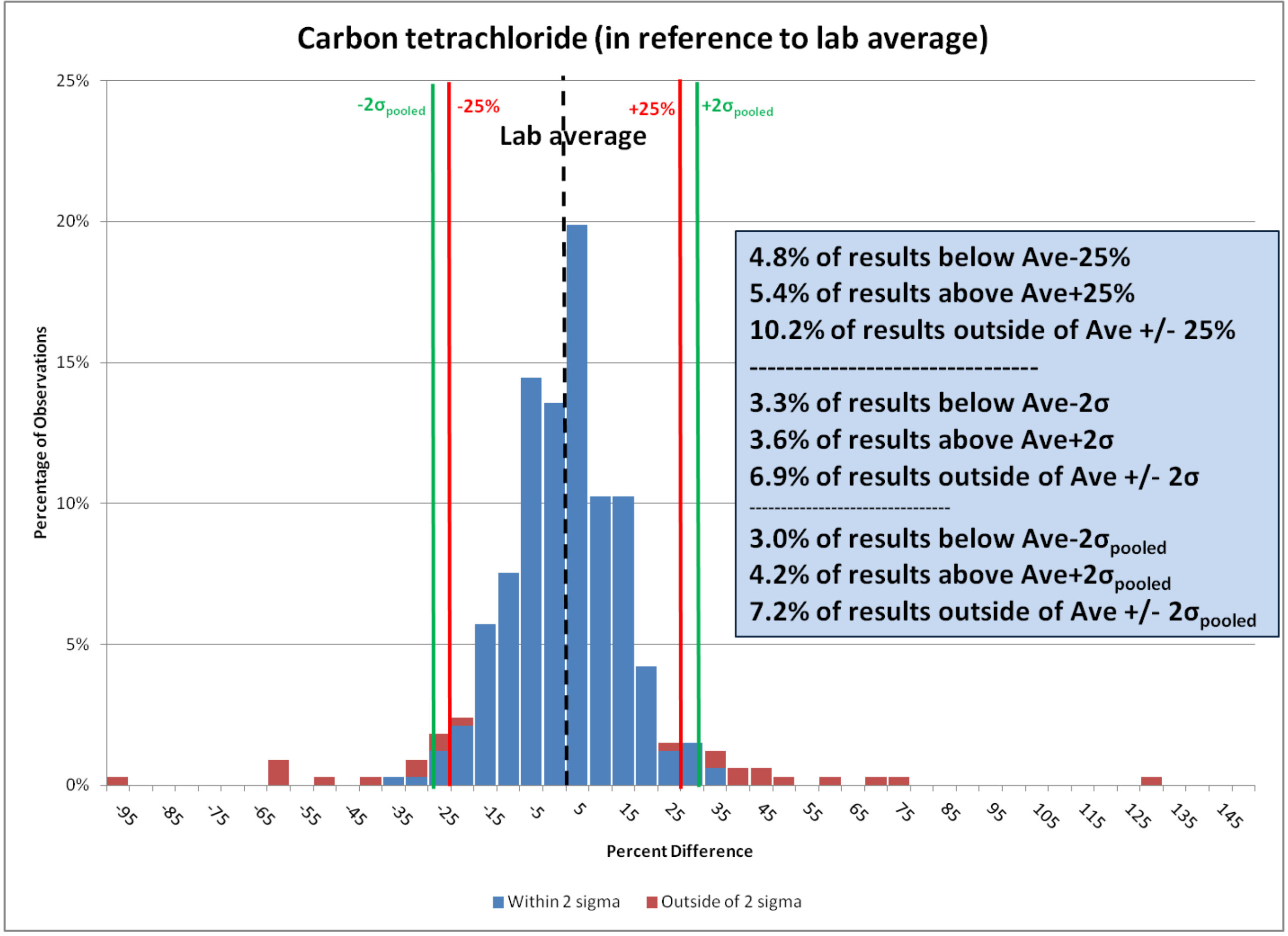
1,2-dichloropropane (in reference to lab average)

2.8% of results below Ave-25%
4.9% of results above Ave+25%
7.7% of results outside of Ave +/- 25%
--------------------------------
2.8% of results below Ave-2σ
4.2% of results above Ave+2σ
7.0% of results outside of Ave +/- 2σ
--------------------------------
2.8% of results below Ave-2$\sigma_{pooled}$
4.9% of results above Ave+2$\sigma_{pooled}$
7.7% of results outside of Ave +/- 2$\sigma_{pooled}$



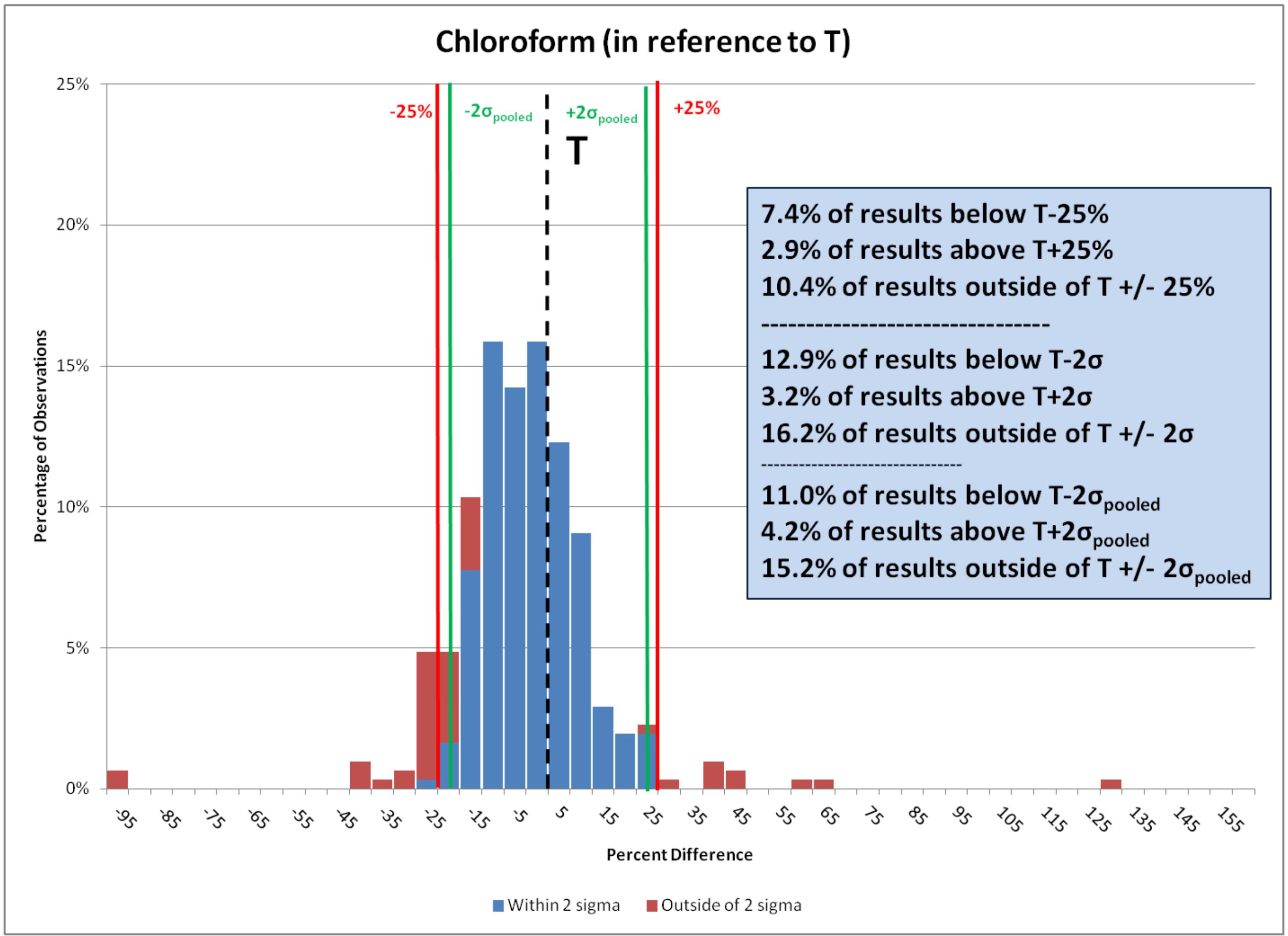
Carbon tetrachloride (in reference to T)

4.5% of results below T-25%
7.2% of results above T+25%
11.7% of results outside of T +/- 25%
--------------------------------
5.4% of results below T-2σ
6.0% of results above T+2σ
11.4% of results outside of T +/- 2σ
--------------------------------
2.7% of results below T-2$\sigma_{pooled}$
5.7% of results above T+2$\sigma_{pooled}$
8.4% of results outside of T +/- 2$\sigma_{pooled}$

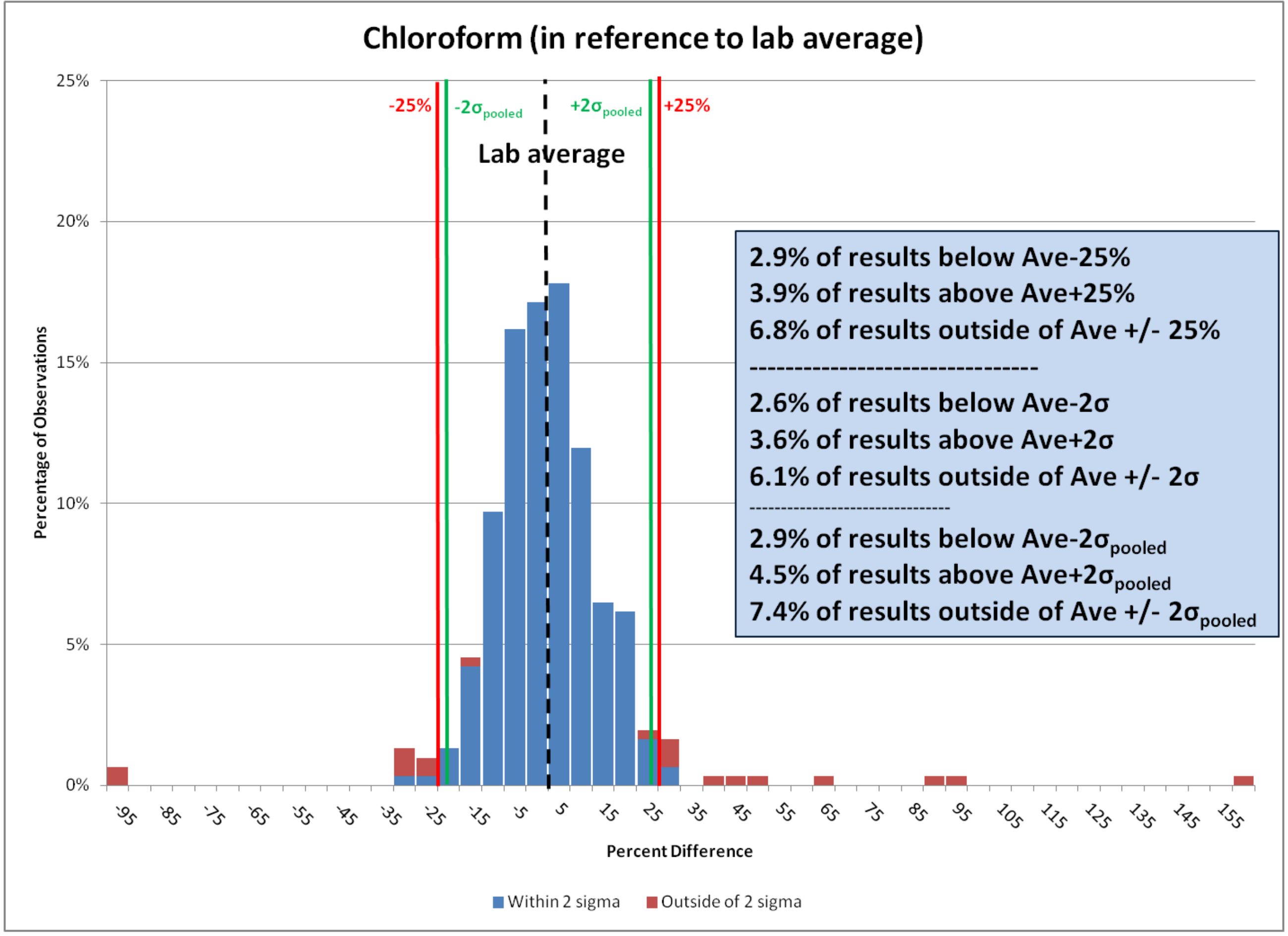


Carbon tetrachloride (in reference to lab average)

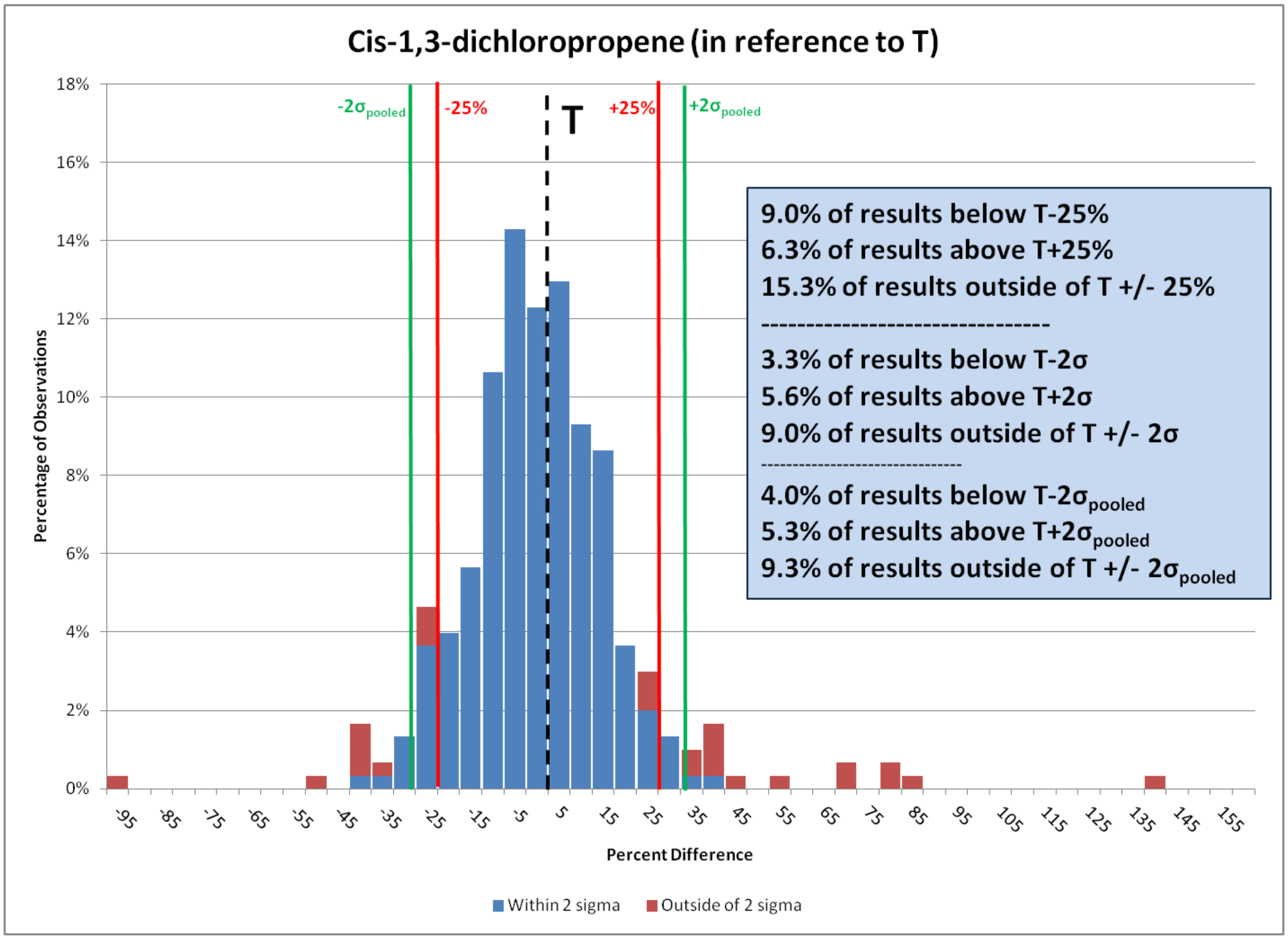
4.8% of results below Ave-25%
5.4% of results above Ave+25%
10.2% of results outside of Ave +/- 25%
--------------------------------
3.3% of results below Ave-2σ
3.6% of results above Ave+2σ
6.9% of results outside of Ave +/- 2σ
--------------------------------
3.0% of results below Ave-2$\sigma_{pooled}$
4.2% of results above Ave+2$\sigma_{pooled}$
7.2% of results outside of Ave +/- 2$\sigma_{pooled}$



7.4% of results below T-25%
2.9% of results above T+25%
10.4% of results outside of T +/- 25%
--------------------------------
12.9% of results below T-2σ
3.2% of results above T+2σ
16.2% of results outside of T +/- 2σ
--------------------------------
11.0% of results below T-$2\sigma_{pooled}$
4.2% of results above T+$2\sigma_{pooled}$
15.2% of results outside of T +/- $2\sigma_{pooled}$

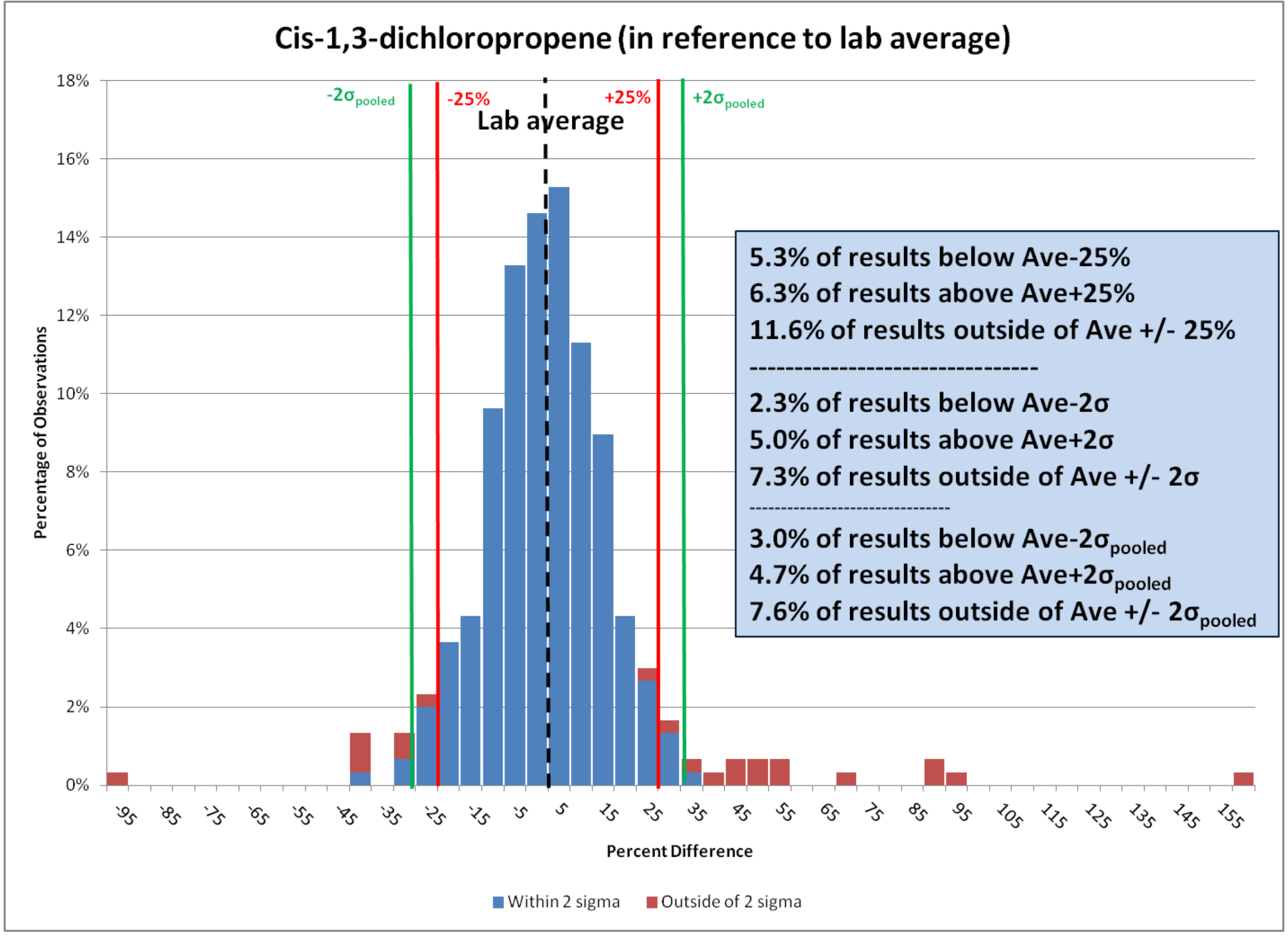


2.9% of results below Ave-25%
3.9% of results above Ave+25%
6.8% of results outside of Ave +/- 25%
--------------------------------
2.6% of results below Ave-2σ
3.6% of results above Ave+2σ
6.1% of results outside of Ave +/- 2σ
--------------------------------
2.9% of results below Ave-$2\sigma_{pooled}$
4.5% of results above Ave+$2\sigma_{pooled}$
7.4% of results outside of Ave +/- $2\sigma_{pooled}$



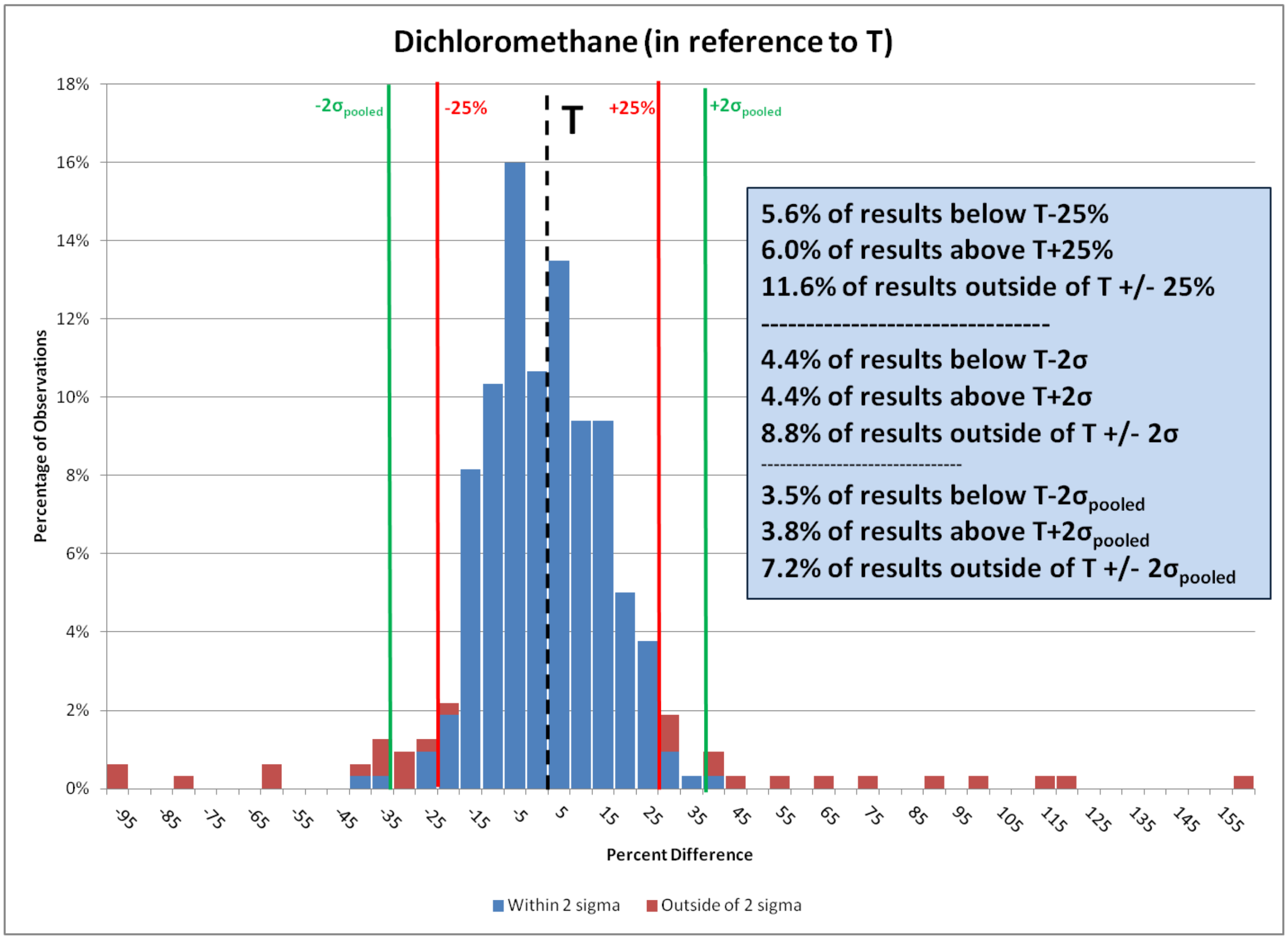
Cis-1,3-dichloropropene (in reference to T)

9.0% of results below T-25%
6.3% of results above T+25%
15.3% of results outside of T +/- 25%
--------------------------------
3.3% of results below T-2σ
5.6% of results above T+2σ
9.0% of results outside of T +/- 2σ
--------------------------------
4.0% of results below $T-2\sigma_{pooled}$
5.3% of results above $T+2\sigma_{pooled}$
9.3% of results outside of $T +/- 2\sigma_{pooled}$



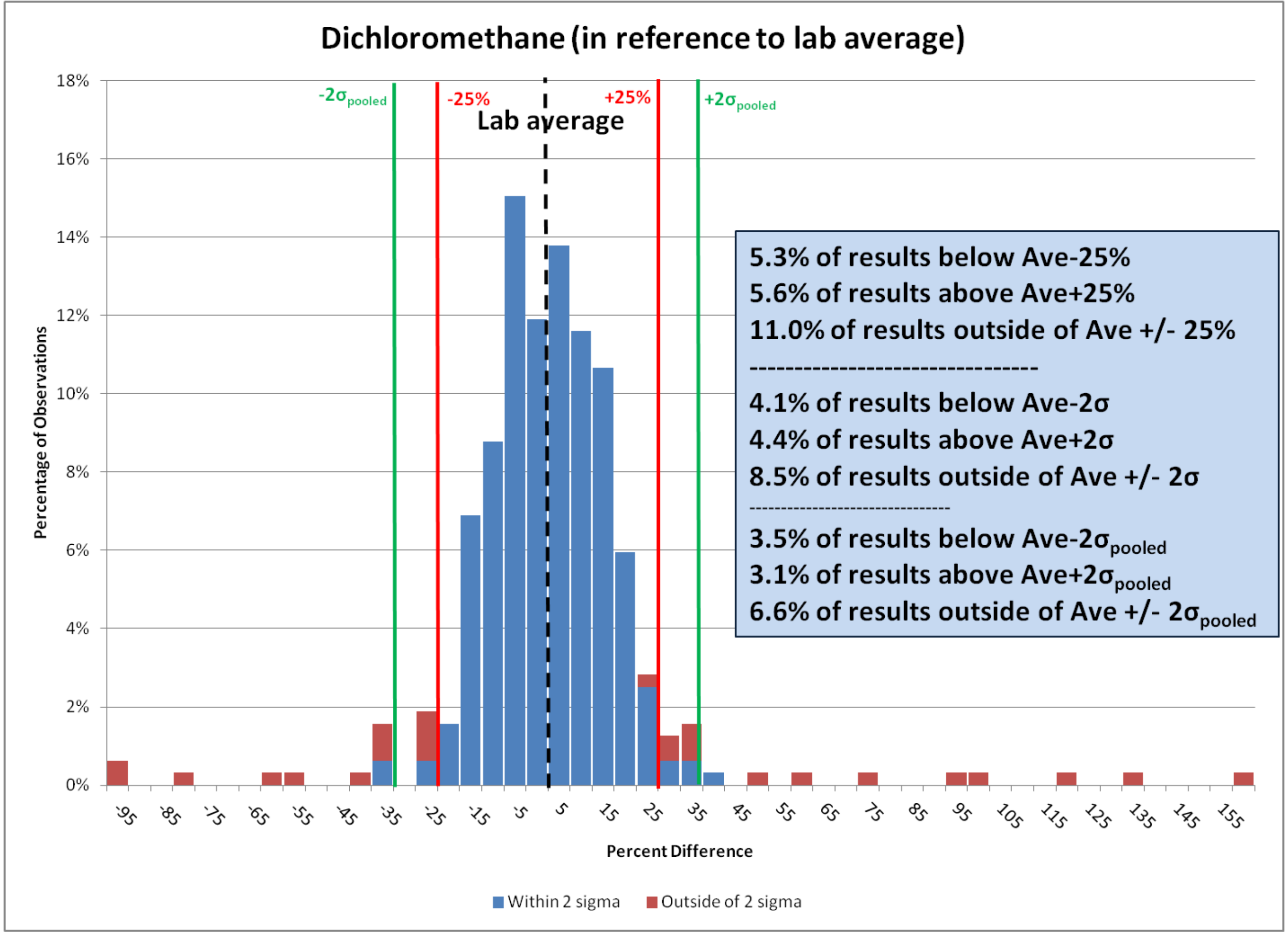
Cis-1,3-dichloropropene (in reference to lab average)

5.3% of results below Ave-25%
6.3% of results above Ave+25%
11.6% of results outside of Ave +/- 25%
--------------------------------
2.3% of results below Ave-2σ
5.0% of results above Ave+2σ
7.3% of results outside of Ave +/- 2σ
--------------------------------
3.0% of results below $Ave-2\sigma_{pooled}$
4.7% of results above $Ave+2\sigma_{pooled}$
7.6% of results outside of $Ave +/- 2\sigma_{pooled}$



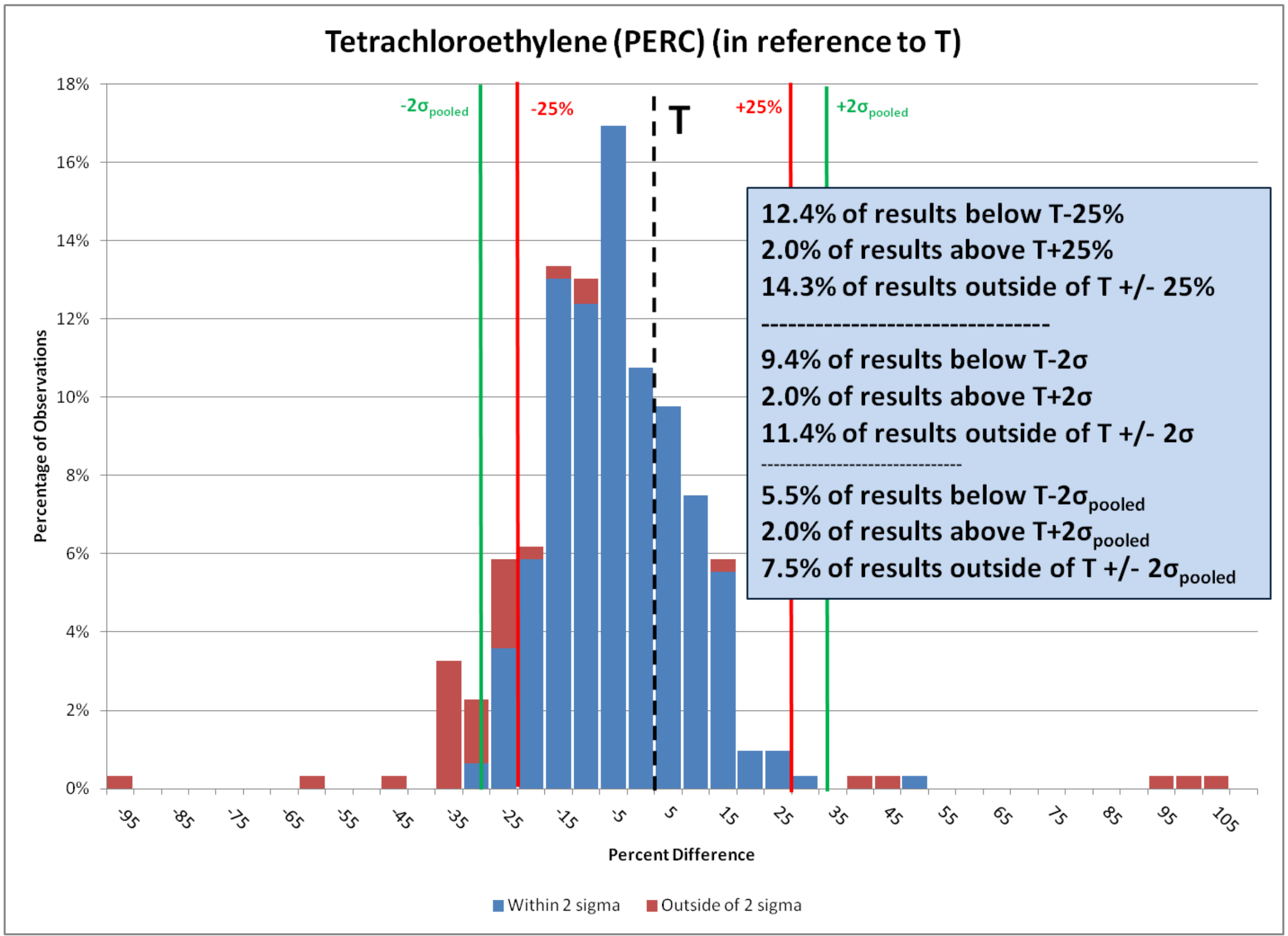
Dichloromethane (in reference to T)

5.6% of results below T-25%
6.0% of results above T+25%
11.6% of results outside of T +/- 25%
--------------------------------
4.4% of results below T-2σ
4.4% of results above T+2σ
8.8% of results outside of T +/- 2σ
--------------------------------
3.5% of results below T-$2\sigma_{pooled}$
3.8% of results above T+$2\sigma_{pooled}$
7.2% of results outside of T +/- $2\sigma_{pooled}$

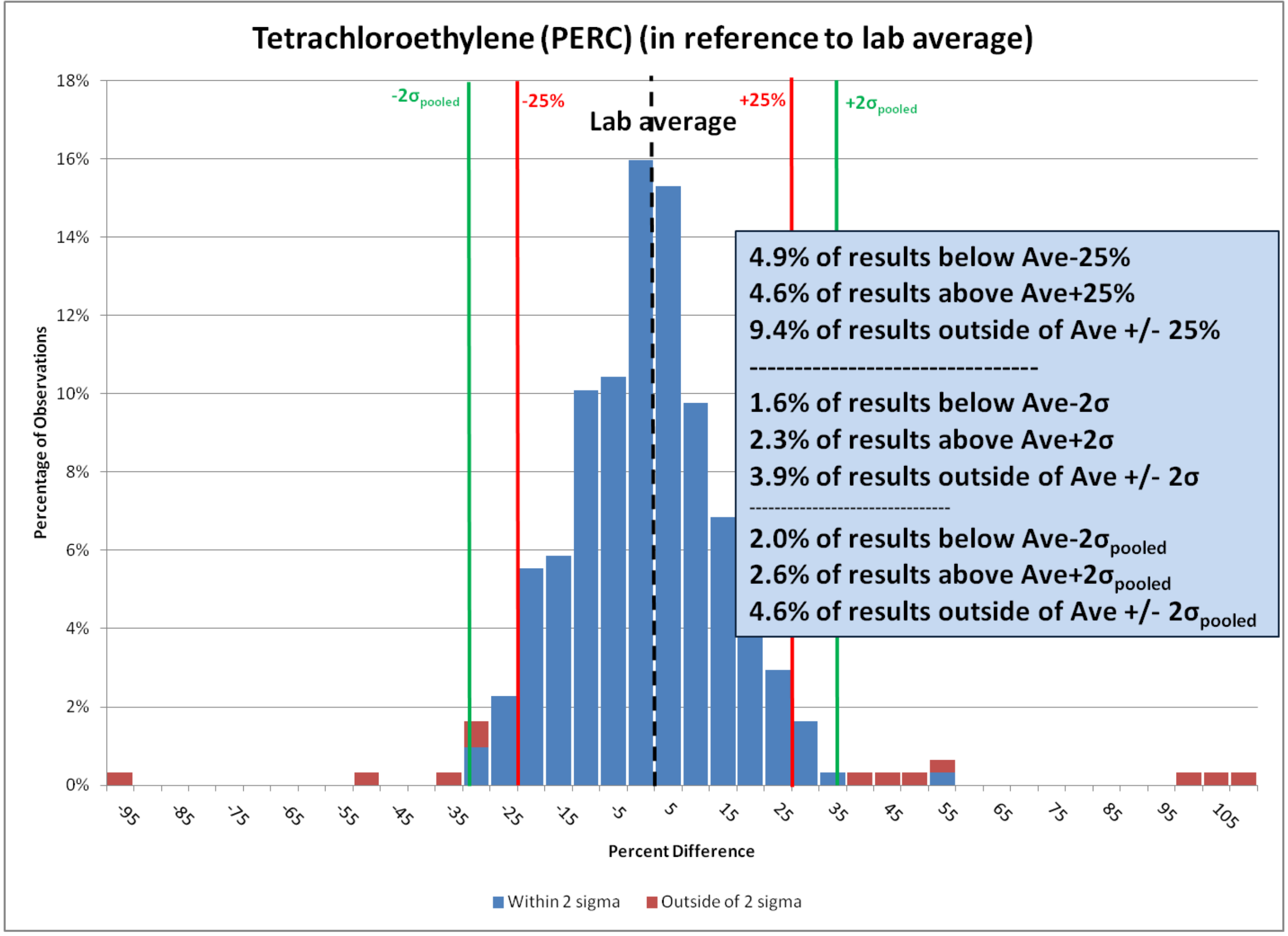


Dichloromethane (in reference to lab average)

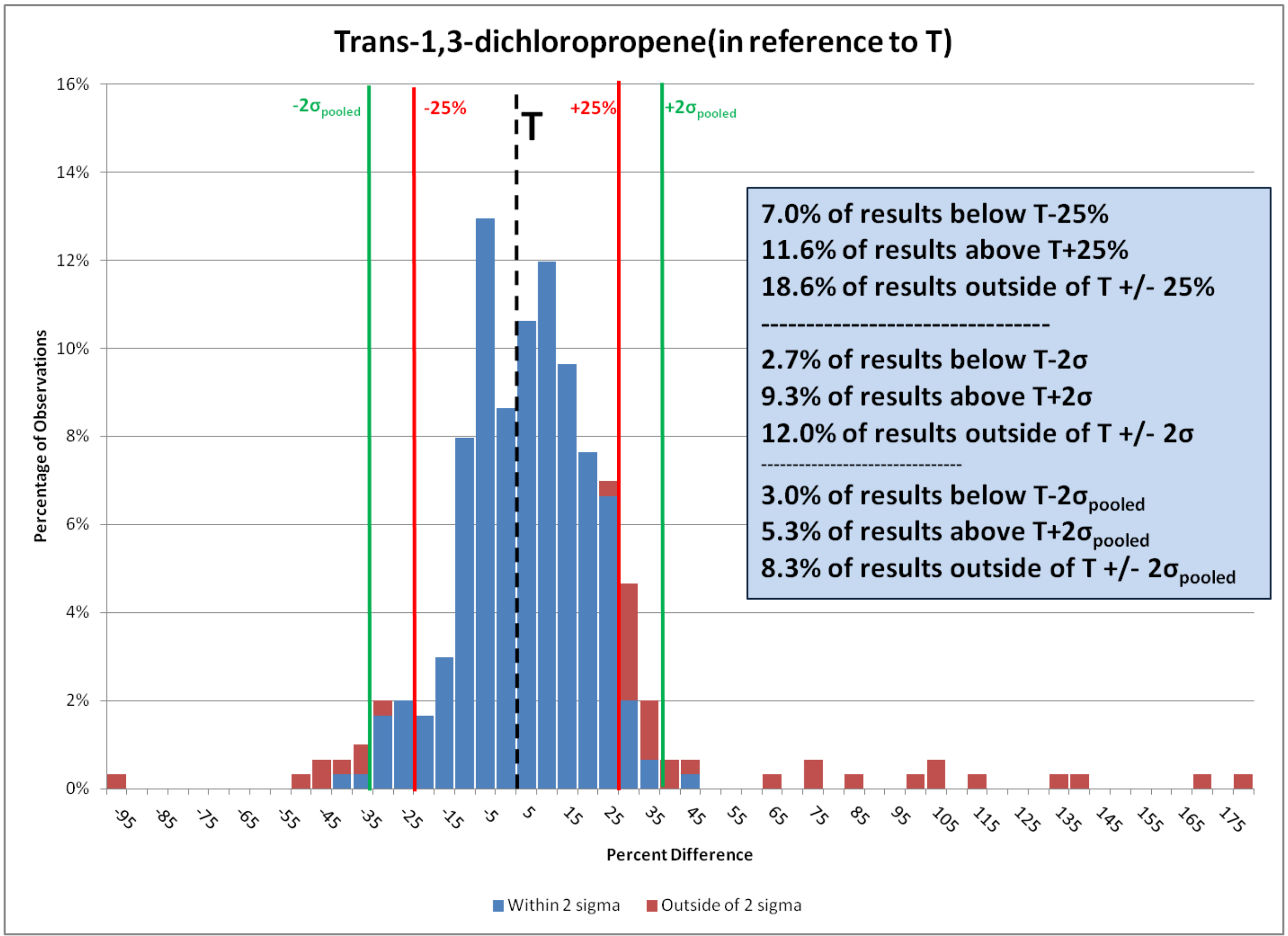
5.3% of results below Ave-25%
5.6% of results above Ave+25%
11.0% of results outside of Ave +/- 25%
--------------------------------
4.1% of results below Ave-2σ
4.4% of results above Ave+2σ
8.5% of results outside of Ave +/- 2σ
--------------------------------
3.5% of results below Ave-$2\sigma_{pooled}$
3.1% of results above Ave+$2\sigma_{pooled}$
6.6% of results outside of Ave +/- $2\sigma_{pooled}$


Tetrachloroethylene (PERC) (in reference to T)

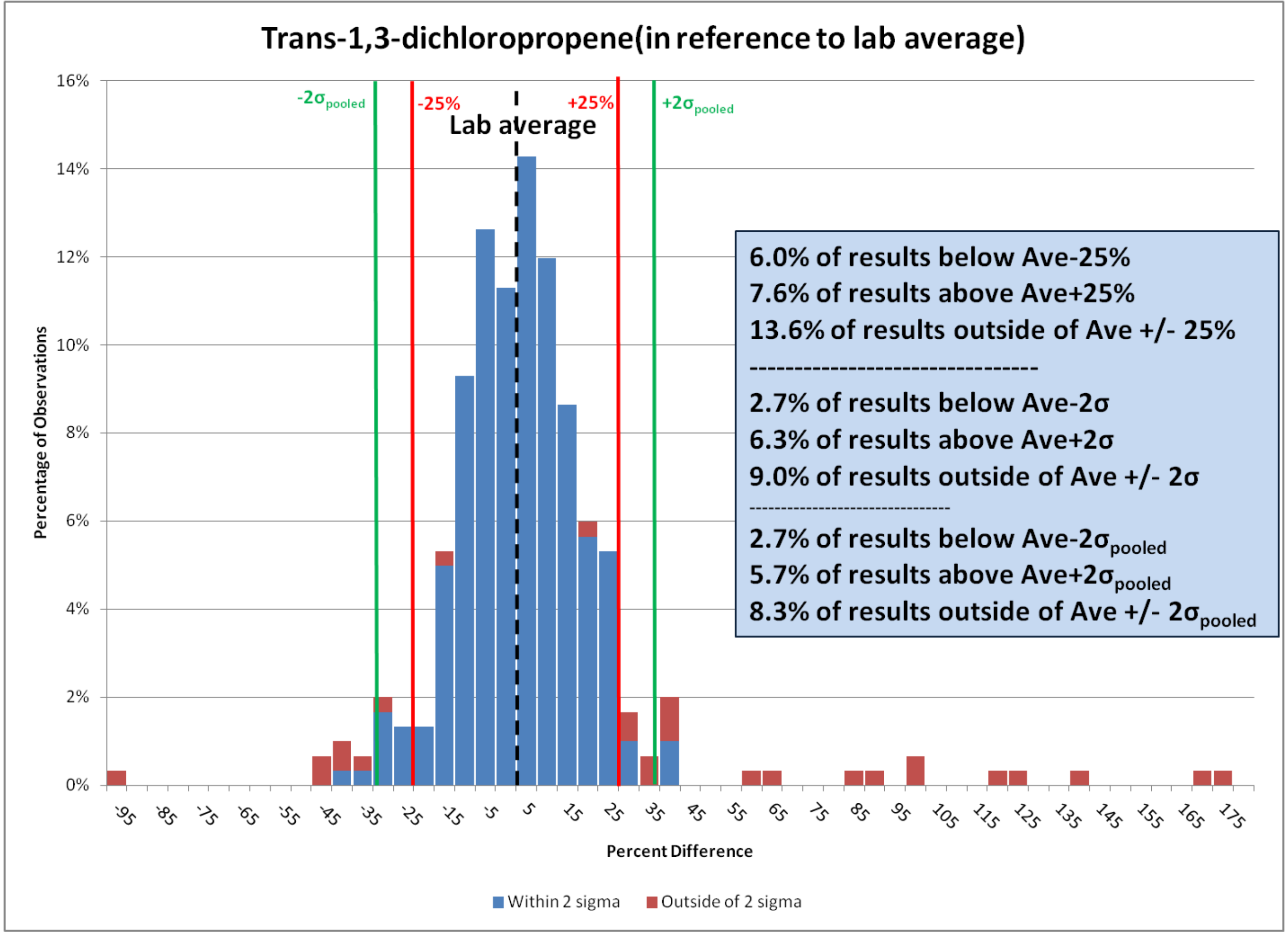
12.4% of results below T-25%
2.0% of results above T+25%
14.3% of results outside of T +/- 25%
--------------------------------
9.4% of results below T-2σ
2.0% of results above T+2σ
11.4% of results outside of T +/- 2σ
--------------------------------
5.5% of results below T-2$\sigma_{pooled}$
2.0% of results above T+2$\sigma_{pooled}$
7.5% of results outside of T +/- 2$\sigma_{pooled}$


Tetrachloroethylene (PERC) (in reference to lab average)

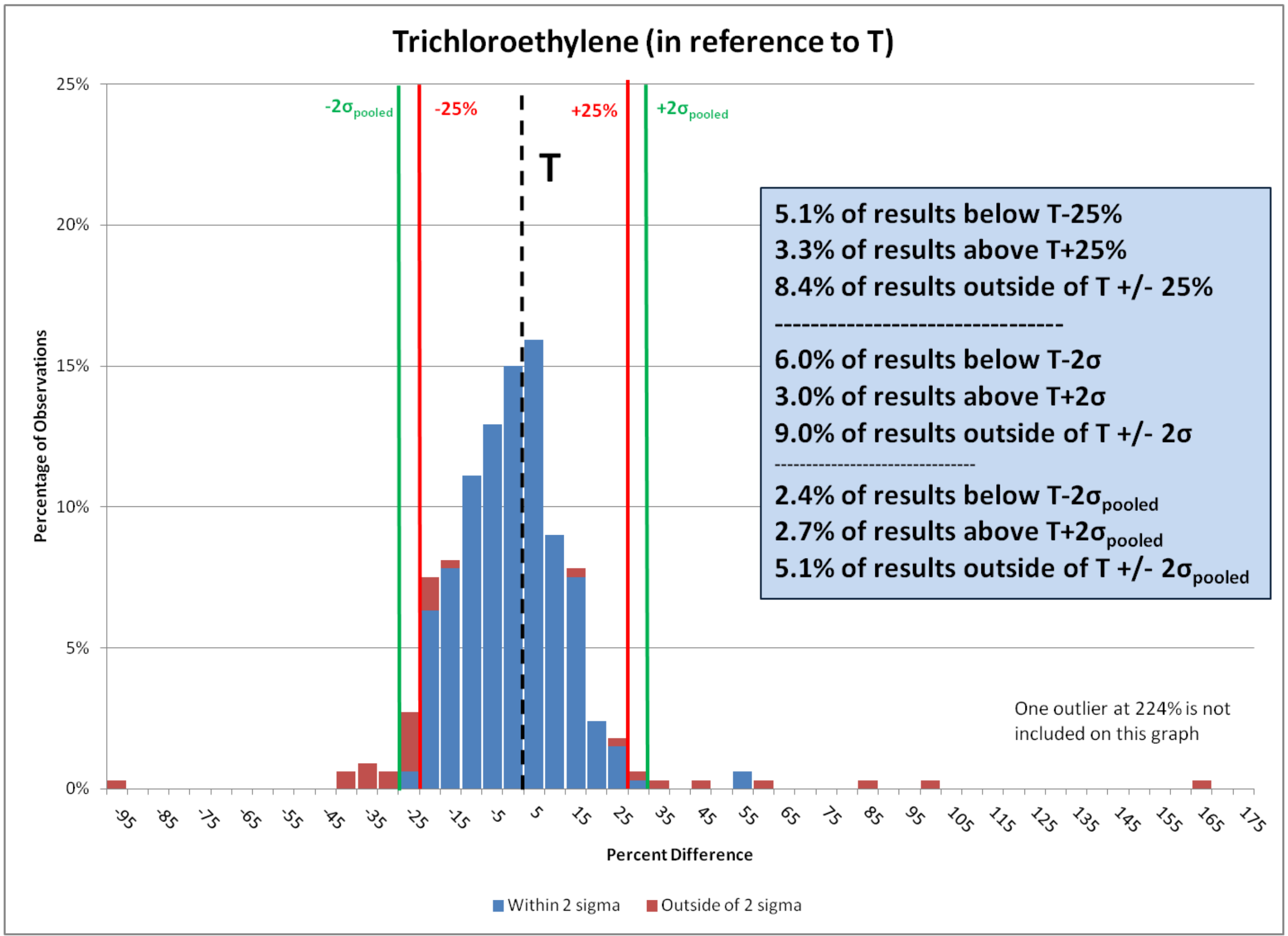
4.9% of results below Ave-25%
4.6% of results above Ave+25%
9.4% of results outside of Ave +/- 25%
--------------------------------
1.6% of results below Ave-2σ
2.3% of results above Ave+2σ
3.9% of results outside of Ave +/- 2σ
--------------------------------
2.0% of results below Ave-2$\sigma_{pooled}$
2.6% of results above Ave+2$\sigma_{pooled}$
4.6% of results outside of Ave +/- 2$\sigma_{pooled}$


Trans-1,3-dichloropropene(in reference to T)

7.0% of results below T-25%
11.6% of results above T+25%
18.6% of results outside of T +/- 25%
--------------------------------
2.7% of results below T-2σ
9.3% of results above T+2σ
12.0% of results outside of T +/- 2σ
--------------------------------
3.0% of results below T-$2\sigma_{pooled}$
5.3% of results above T+$2\sigma_{pooled}$
8.3% of results outside of T +/- $2\sigma_{pooled}$


Trans-1,3-dichloropropene(in reference to lab average)

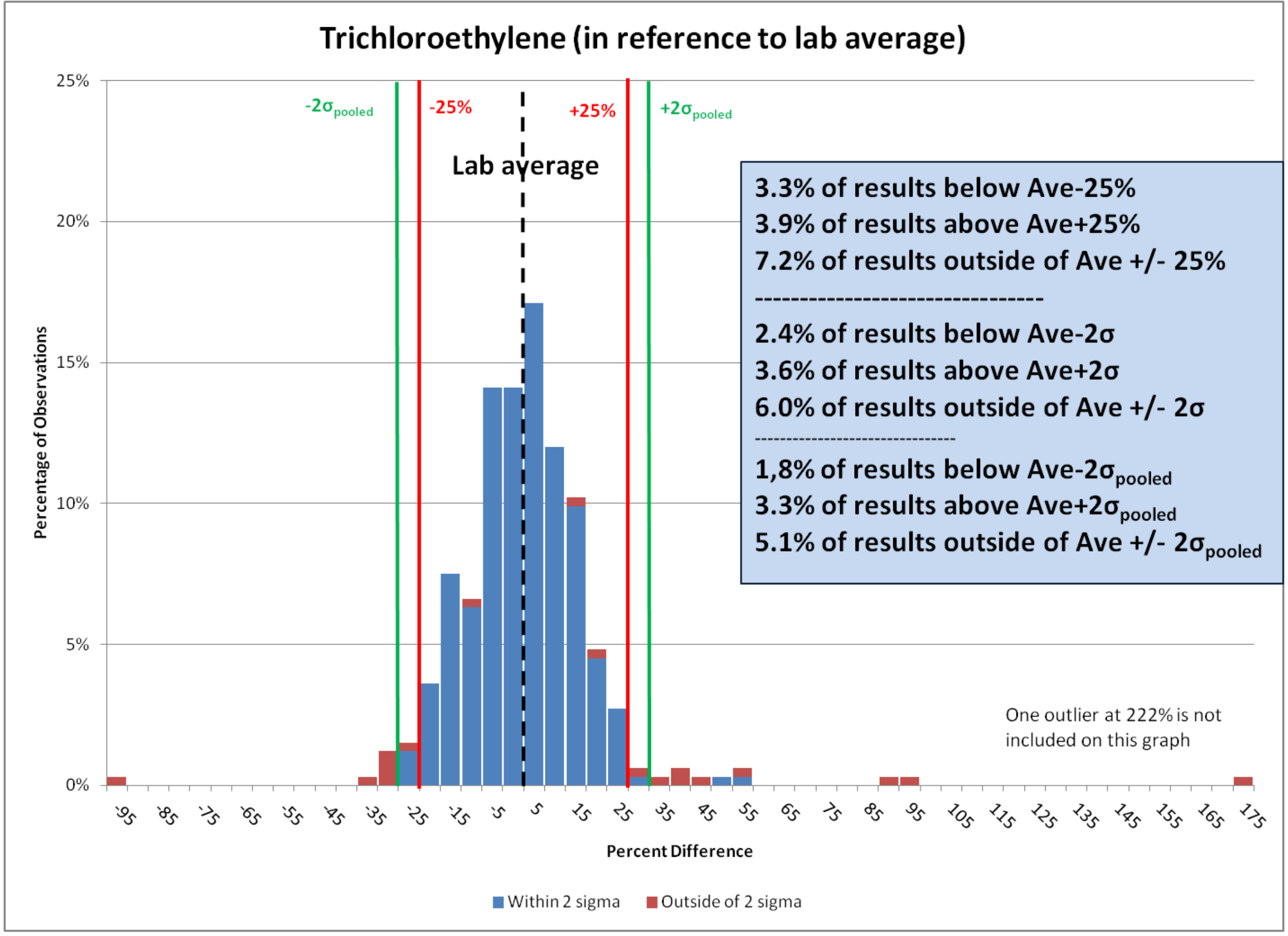
6.0% of results below Ave-25%
7.6% of results above Ave+25%
13.6% of results outside of Ave +/- 25%
--------------------------------
2.7% of results below Ave-2σ
6.3% of results above Ave+2σ
9.0% of results outside of Ave +/- 2σ
--------------------------------
2.7% of results below Ave-$2\sigma_{pooled}$
5.7% of results above Ave+$2\sigma_{pooled}$
8.3% of results outside of Ave +/- $2\sigma_{pooled}$



## Trichloroethylene (in reference to T)

5.1% of results below T-25%
3.3% of results above T+25%
8.4% of results outside of T +/- 25%
--------------------------------
6.0% of results below T-2σ
3.0% of results above T+2σ
9.0% of results outside of T +/- 2σ
--------------------------------
2.4% of results below T-$2\sigma_{pooled}$
2.7% of results above T+$2\sigma_{pooled}$
5.1% of results outside of T +/- $2\sigma_{pooled}$

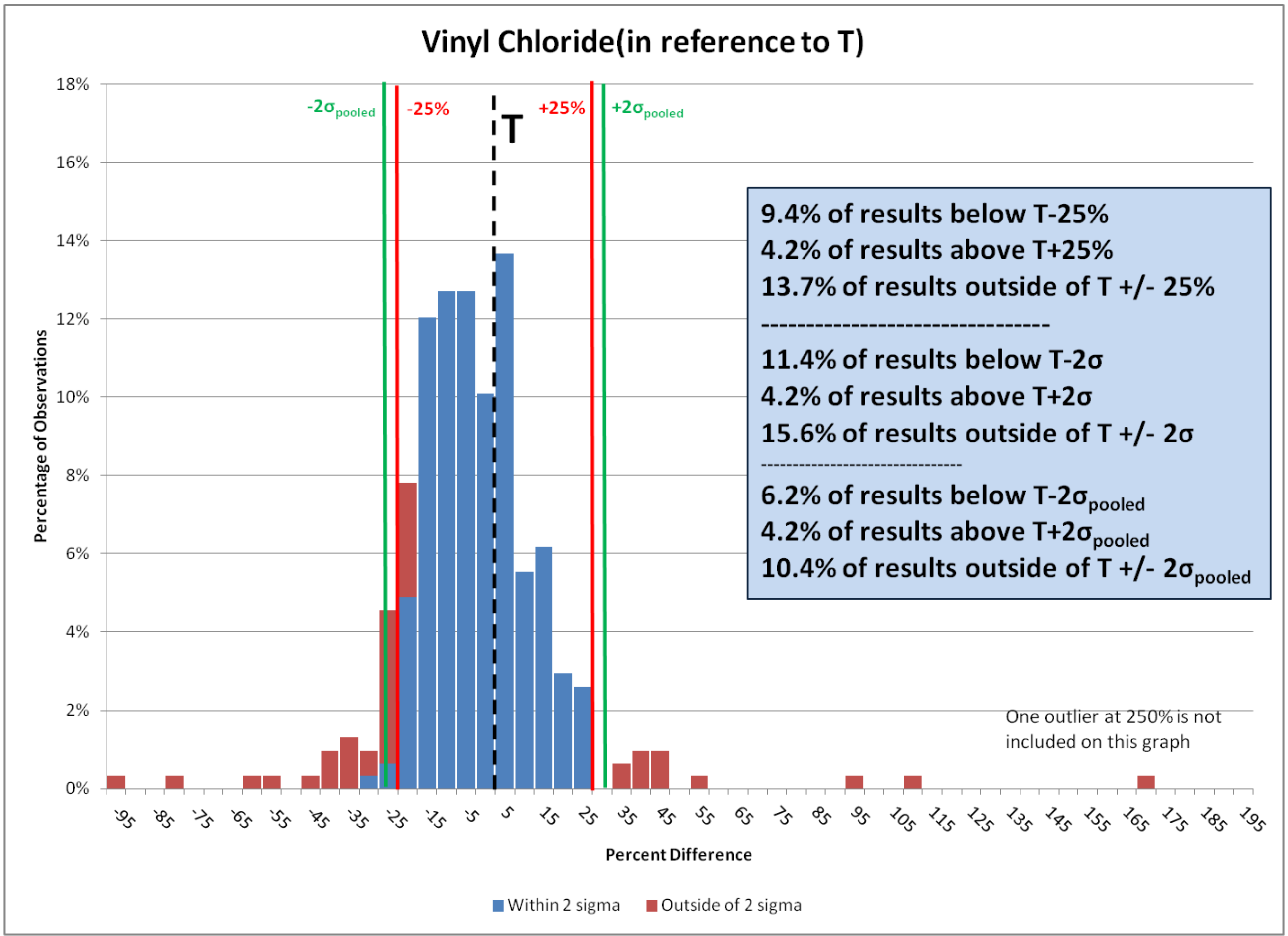
One outlier at 224% is not included on this graph



## Trichloroethylene (in reference to lab average)

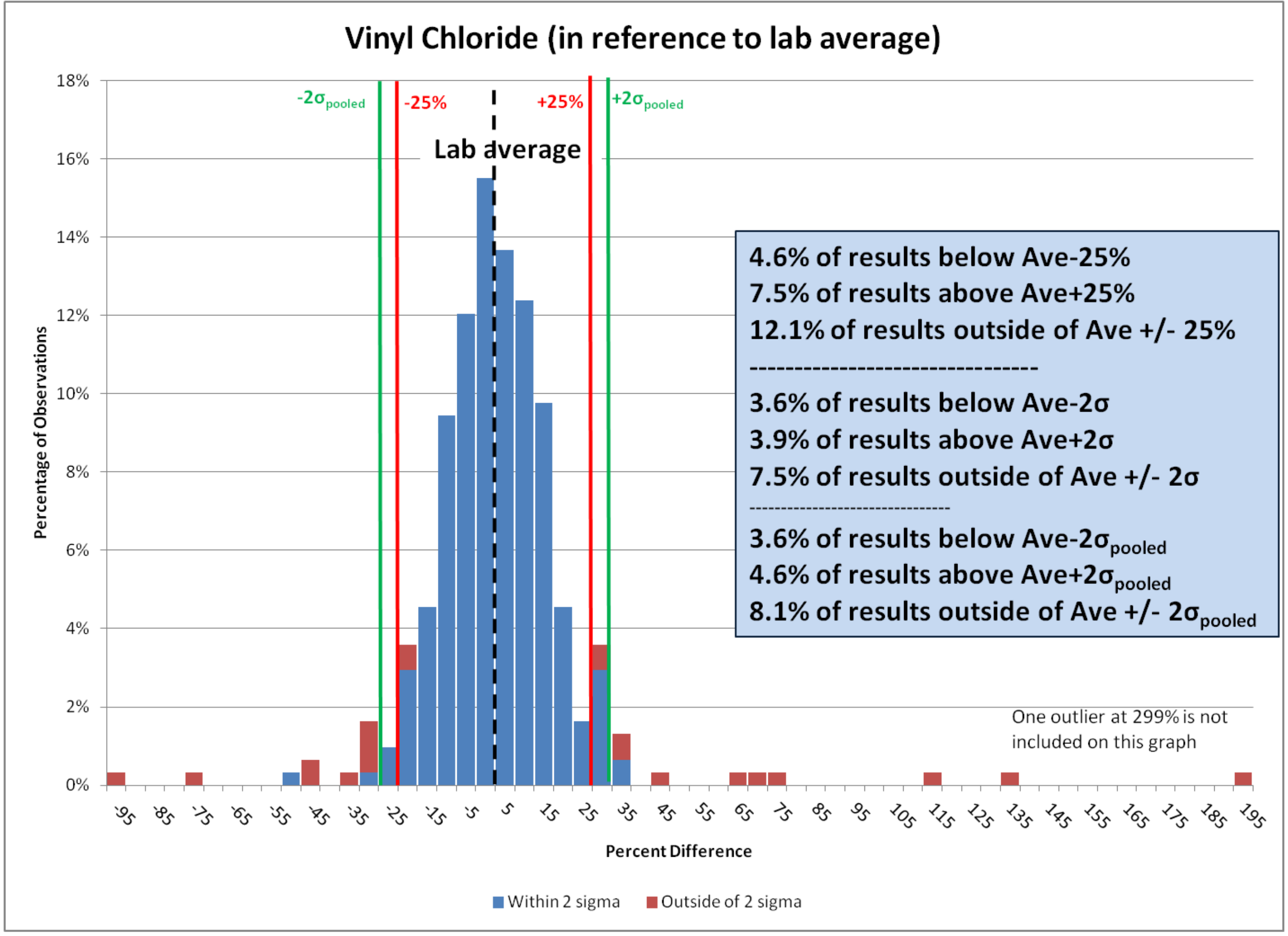
3.3% of results below Ave-25%
3.9% of results above Ave+25%
7.2% of results outside of Ave +/- 25%
--------------------------------
2.4% of results below Ave-2σ
3.6% of results above Ave+2σ
6.0% of results outside of Ave +/- 2σ
--------------------------------
1,8% of results below Ave-$2\sigma_{pooled}$
3.3% of results above Ave+$2\sigma_{pooled}$
5.1% of results outside of Ave +/- $2\sigma_{pooled}$

One outlier at 222% is not included on this graph



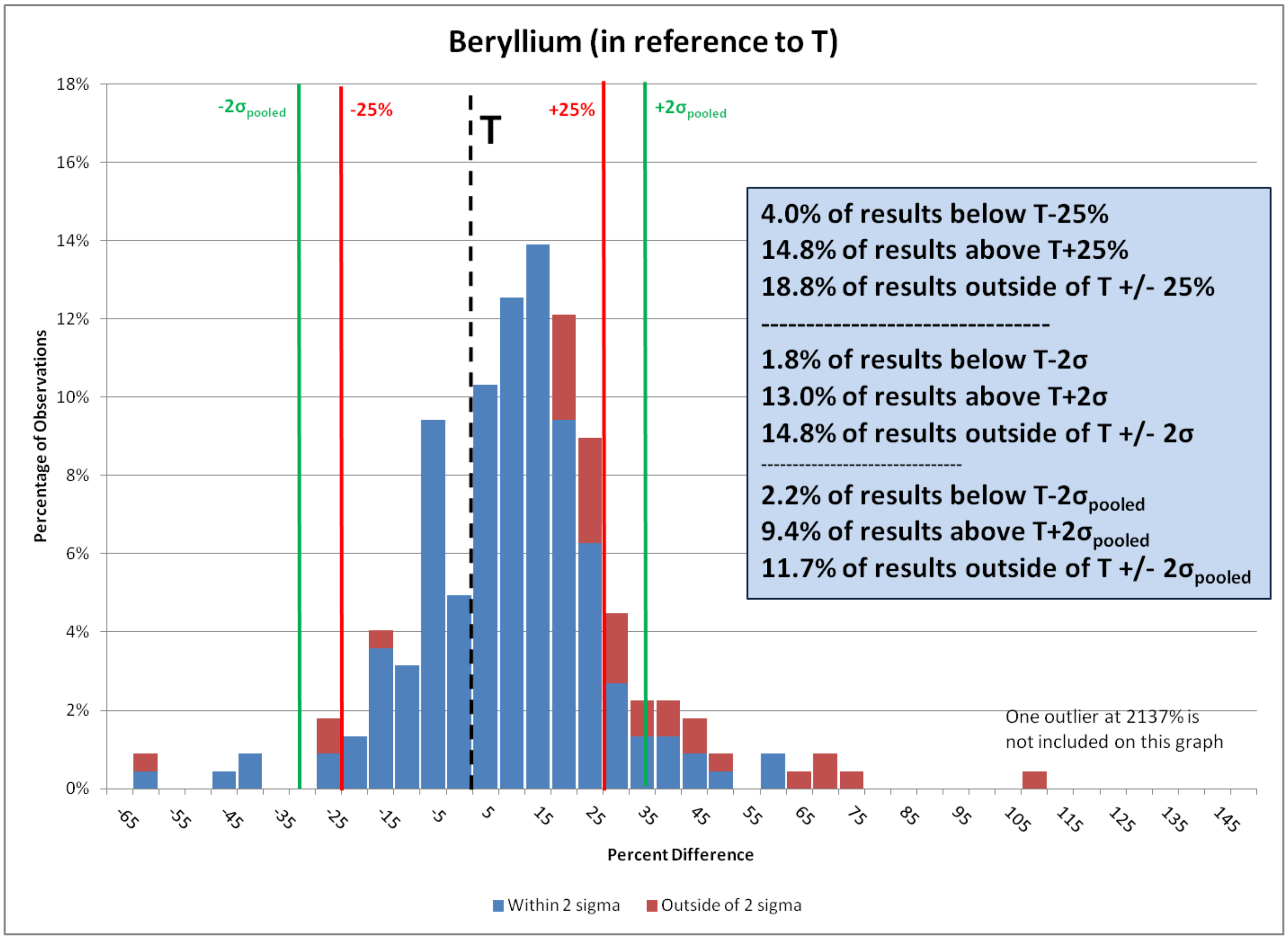
9.4% of results below T-25%
4.2% of results above T+25%
13.7% of results outside of T +/- 25%
--------------------------------
11.4% of results below T-2σ
4.2% of results above T+2σ
15.6% of results outside of T +/- 2σ
--------------------------------
6.2% of results below T-2$\sigma_{pooled}$
4.2% of results above T+2$\sigma_{pooled}$
10.4% of results outside of T +/- 2$\sigma_{pooled}$

One outlier at 250% is not included on this graph

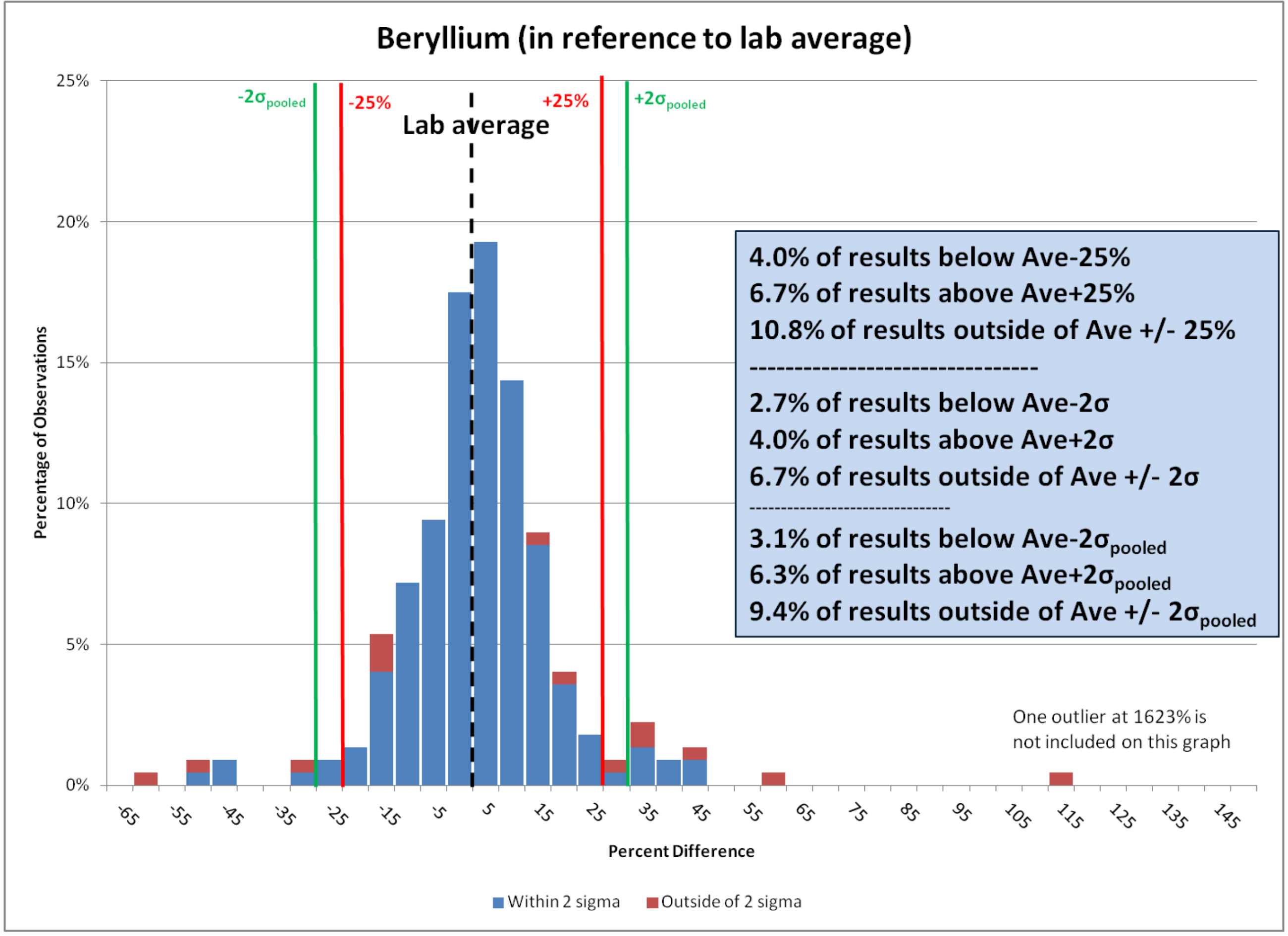


4.6% of results below Ave-25%
7.5% of results above Ave+25%
12.1% of results outside of Ave +/- 25%
--------------------------------
3.6% of results below Ave-2σ
3.9% of results above Ave+2σ
7.5% of results outside of Ave +/- 2σ
--------------------------------
3.6% of results below Ave-2$\sigma_{pooled}$
4.6% of results above Ave+2$\sigma_{pooled}$
8.1% of results outside of Ave +/- 2$\sigma_{pooled}$

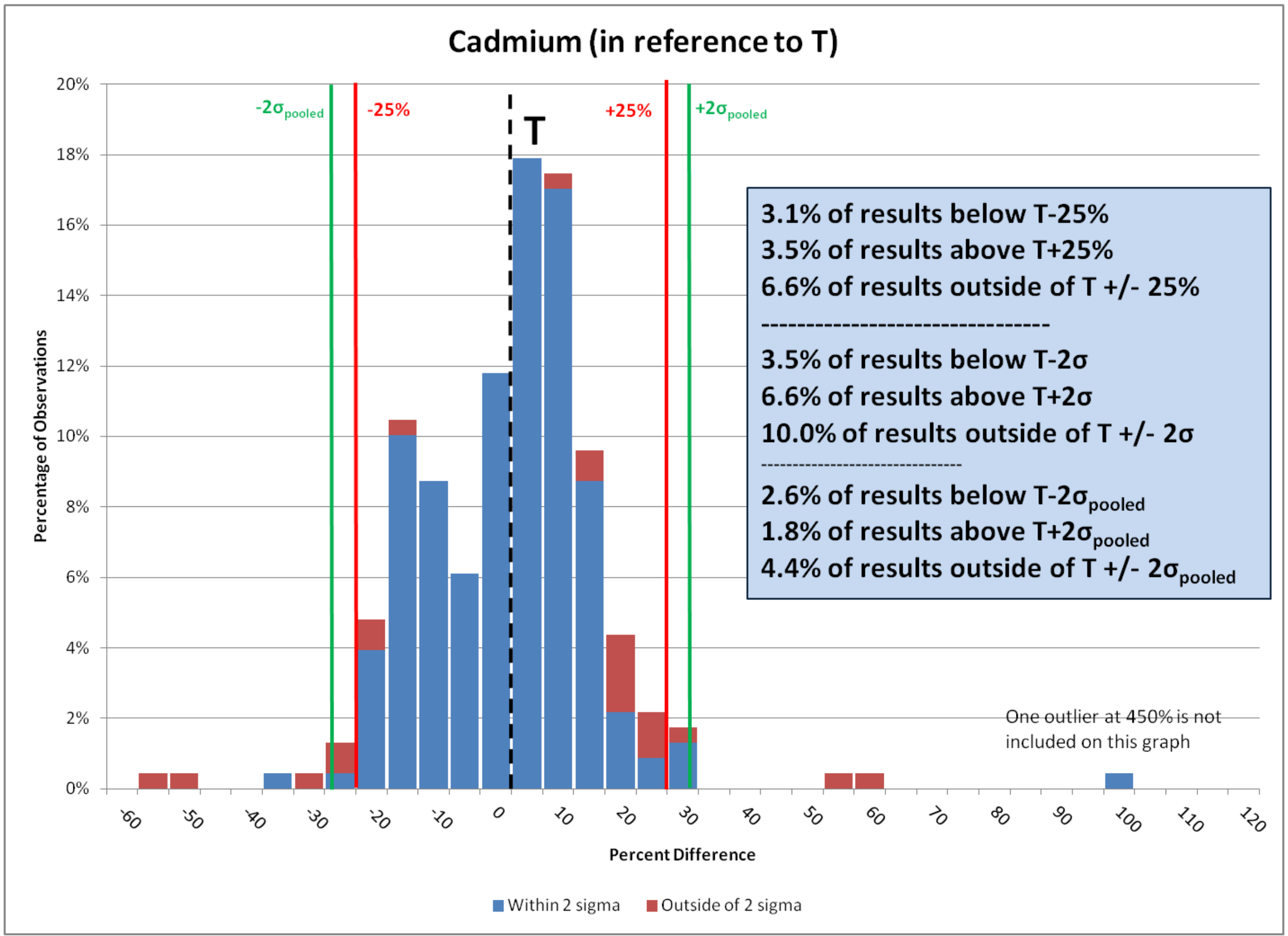
One outlier at 299% is not included on this graph



4.0% of results below T-25%
14.8% of results above T+25%
18.8% of results outside of T +/- 25%
--------------------------------
1.8% of results below T-2σ
13.0% of results above T+2σ
14.8% of results outside of T +/- 2σ
--------------------------------
2.2% of results below T-2$σ_{pooled}$
9.4% of results above T+2$σ_{pooled}$
11.7% of results outside of T +/- 2$σ_{pooled}$

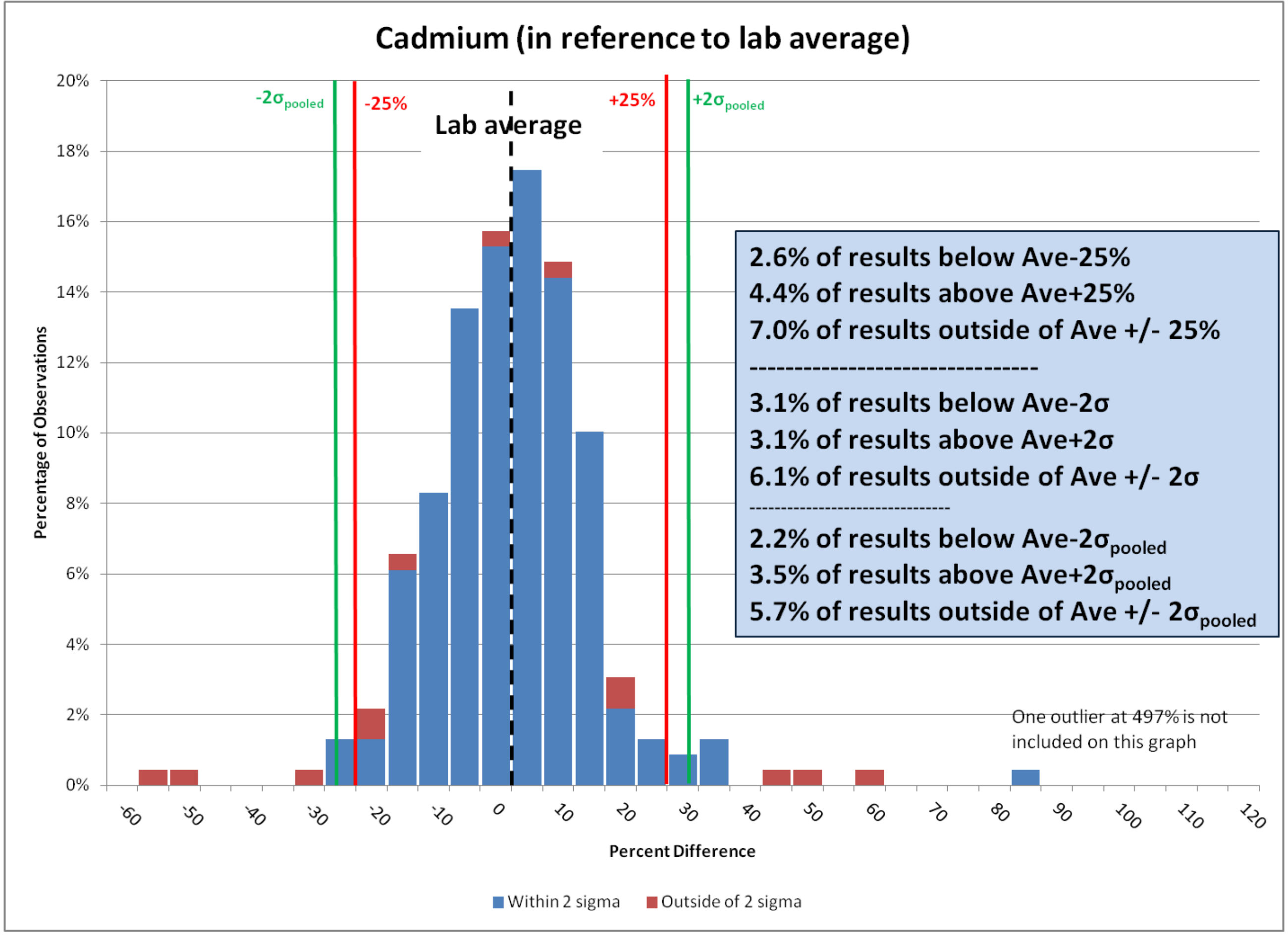


4.0% of results below Ave-25%
6.7% of results above Ave+25%
10.8% of results outside of Ave +/- 25%
--------------------------------
2.7% of results below Ave-2σ
4.0% of results above Ave+2σ
6.7% of results outside of Ave +/- 2σ
--------------------------------
3.1% of results below Ave-2$σ_{pooled}$
6.3% of results above Ave+2$σ_{pooled}$
9.4% of results outside of Ave +/- 2$σ_{pooled}$


Cadmium (in reference to T)

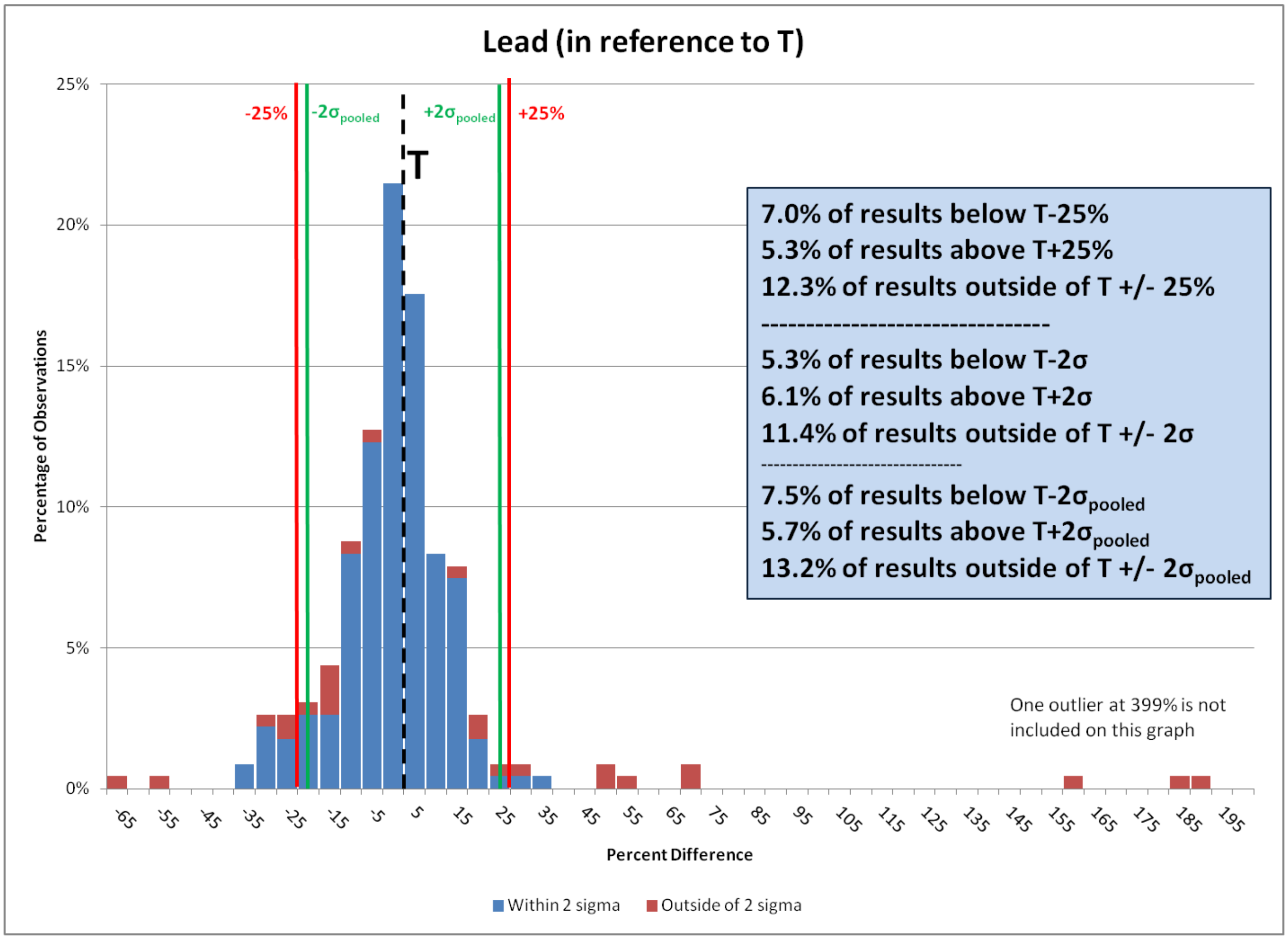
3.1% of results below T-25%
3.5% of results above T+25%
6.6% of results outside of T +/- 25%
--------------------------------
3.5% of results below T-2σ
6.6% of results above T+2σ
10.0% of results outside of T +/- 2σ
--------------------------------
2.6% of results below T-$2\sigma_{pooled}$
1.8% of results above T+$2\sigma_{pooled}$
4.4% of results outside of T +/- $2\sigma_{pooled}$

One outlier at 450% is not included on this graph


Cadmium (in reference to lab average)

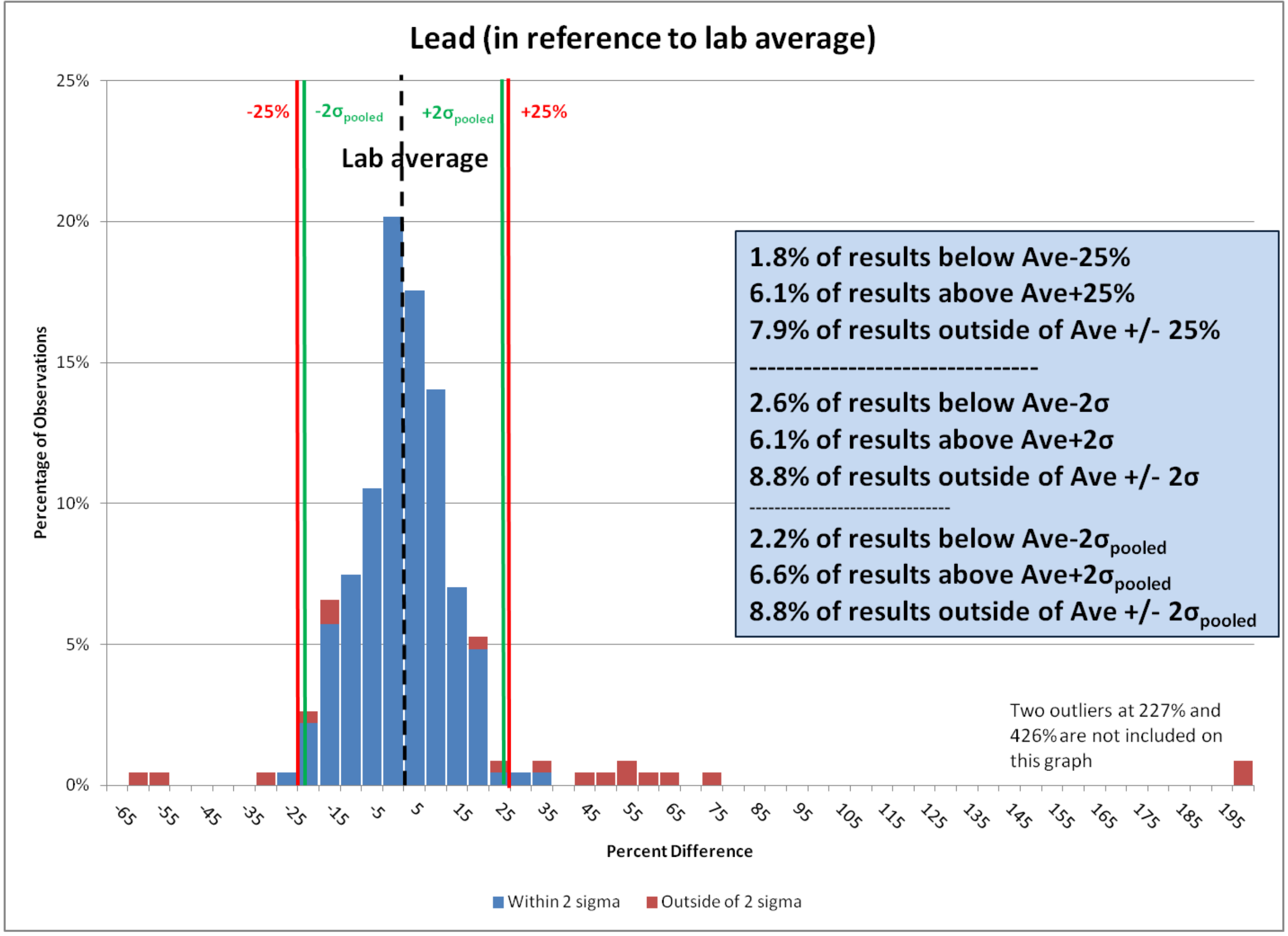
2.6% of results below Ave-25%
4.4% of results above Ave+25%
7.0% of results outside of Ave +/- 25%
--------------------------------
3.1% of results below Ave-2σ
3.1% of results above Ave+2σ
6.1% of results outside of Ave +/- 2σ
--------------------------------
2.2% of results below Ave-$2\sigma_{pooled}$
3.5% of results above Ave+$2\sigma_{pooled}$
5.7% of results outside of Ave +/- $2\sigma_{pooled}$

One outlier at 497% is not included on this graph


Lead (in reference to T)

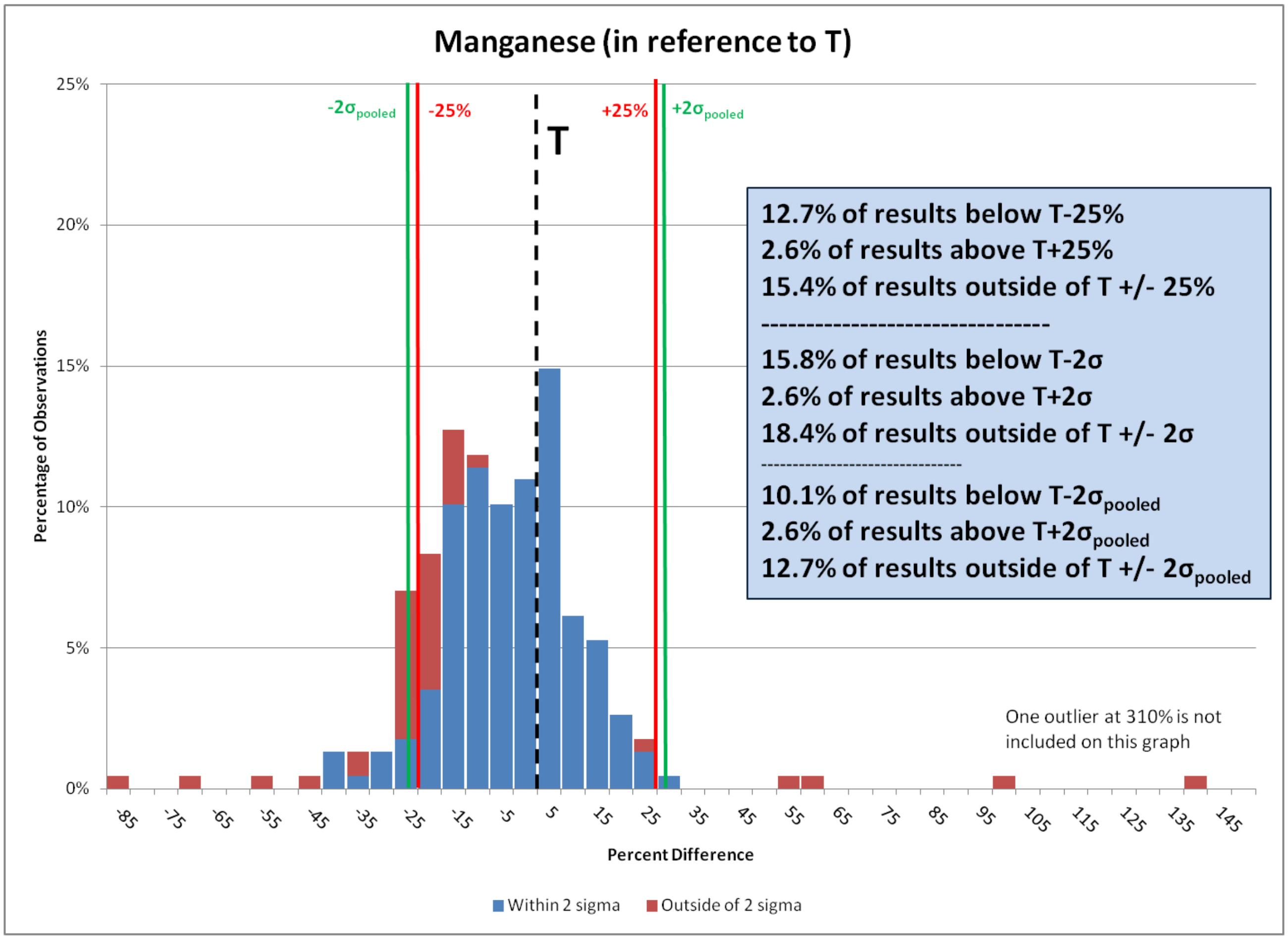
7.0% of results below T-25%
5.3% of results above T+25%
12.3% of results outside of T +/- 25%
--------------------------------
5.3% of results below T-2σ
6.1% of results above T+2σ
11.4% of results outside of T +/- 2σ
--------------------------------
7.5% of results below T-2$\sigma_{pooled}$
5.7% of results above T+2$\sigma_{pooled}$
13.2% of results outside of T +/- 2$\sigma_{pooled}$

One outlier at 399% is not included on this graph


Lead (in reference to lab average)

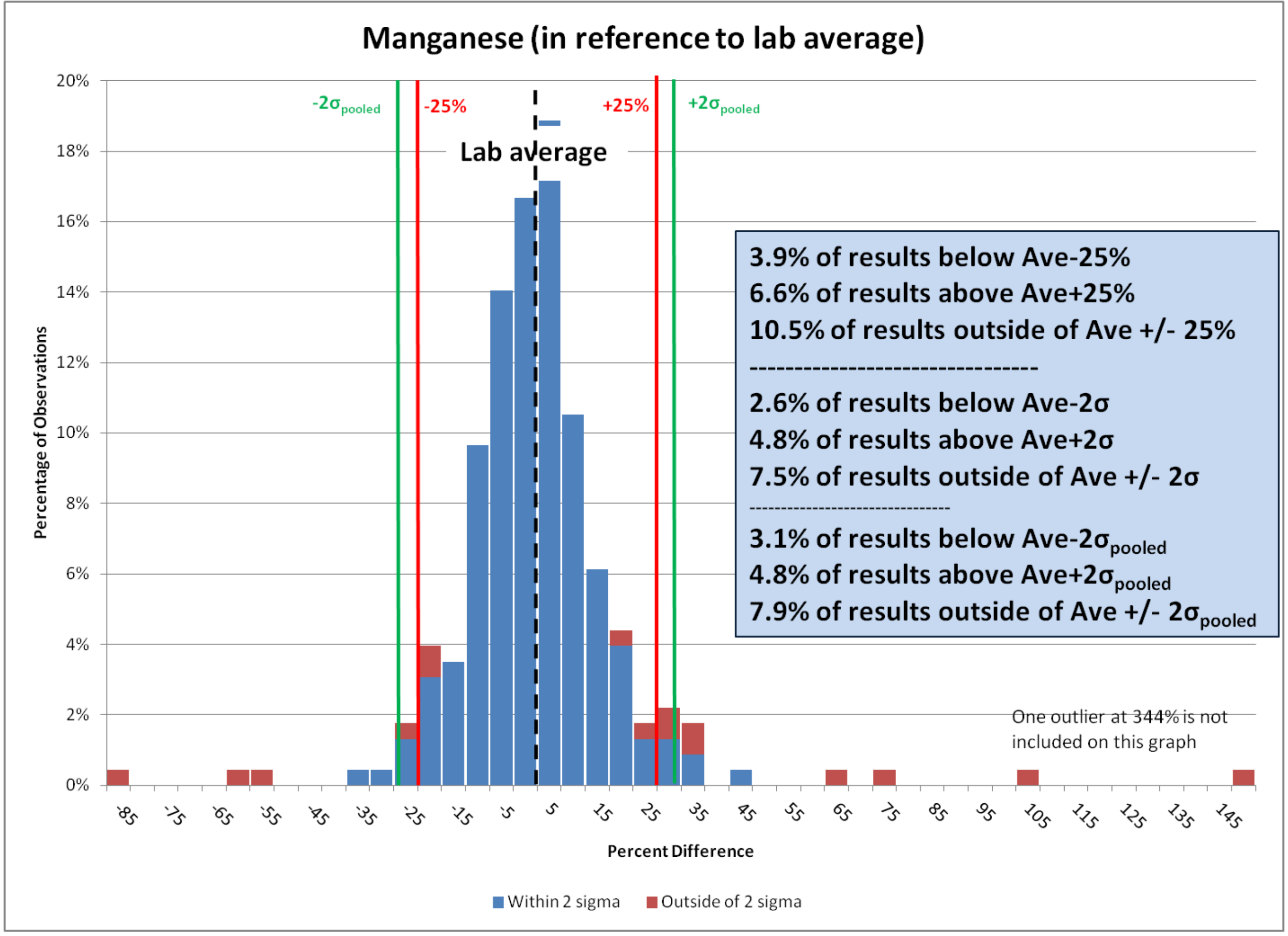
1.8% of results below Ave-25%
6.1% of results above Ave+25%
7.9% of results outside of Ave +/- 25%
--------------------------------
2.6% of results below Ave-2σ
6.1% of results above Ave+2σ
8.8% of results outside of Ave +/- 2σ
--------------------------------
2.2% of results below Ave-2$\sigma_{pooled}$
6.6% of results above Ave+2$\sigma_{pooled}$
8.8% of results outside of Ave +/- 2$\sigma_{pooled}$

Two outliers at 227% and 426% are not included on this graph



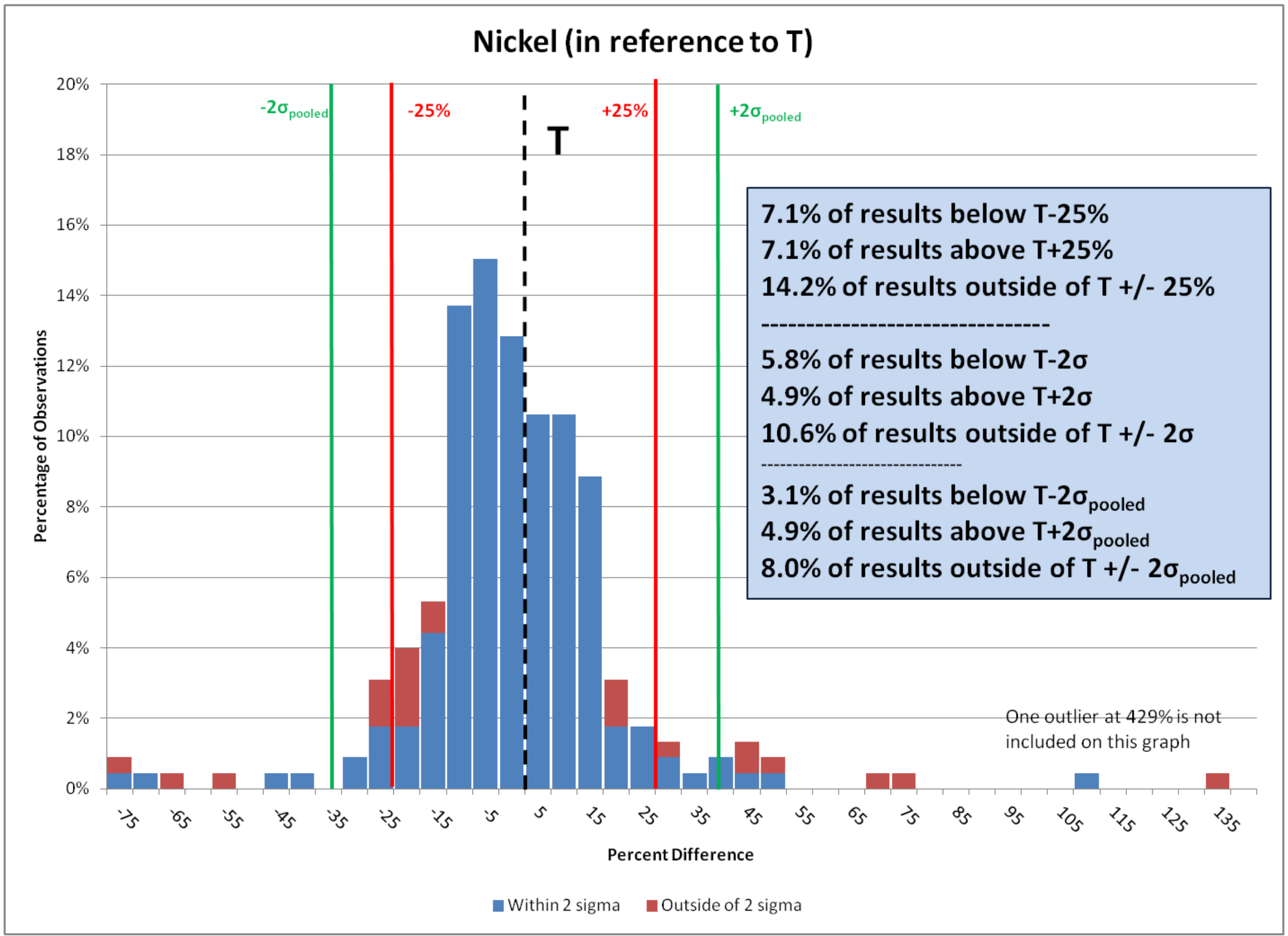
12.7% of results below T-25%
2.6% of results above T+25%
15.4% of results outside of T +/- 25%
--------------------------------
15.8% of results below T-2σ
2.6% of results above T+2σ
18.4% of results outside of T +/- 2σ
--------------------------------
10.1% of results below T-$2\sigma_{pooled}$
2.6% of results above T+$2\sigma_{pooled}$
12.7% of results outside of T +/- $2\sigma_{pooled}$

One outlier at 310% is not included on this graph



3.9% of results below Ave-25%
6.6% of results above Ave+25%
10.5% of results outside of Ave +/- 25%
--------------------------------
2.6% of results below Ave-2σ
4.8% of results above Ave+2σ
7.5% of results outside of Ave +/- 2σ
--------------------------------
3.1% of results below Ave-$2\sigma_{pooled}$
4.8% of results above Ave+$2\sigma_{pooled}$
7.9% of results outside of Ave +/- $2\sigma_{pooled}$

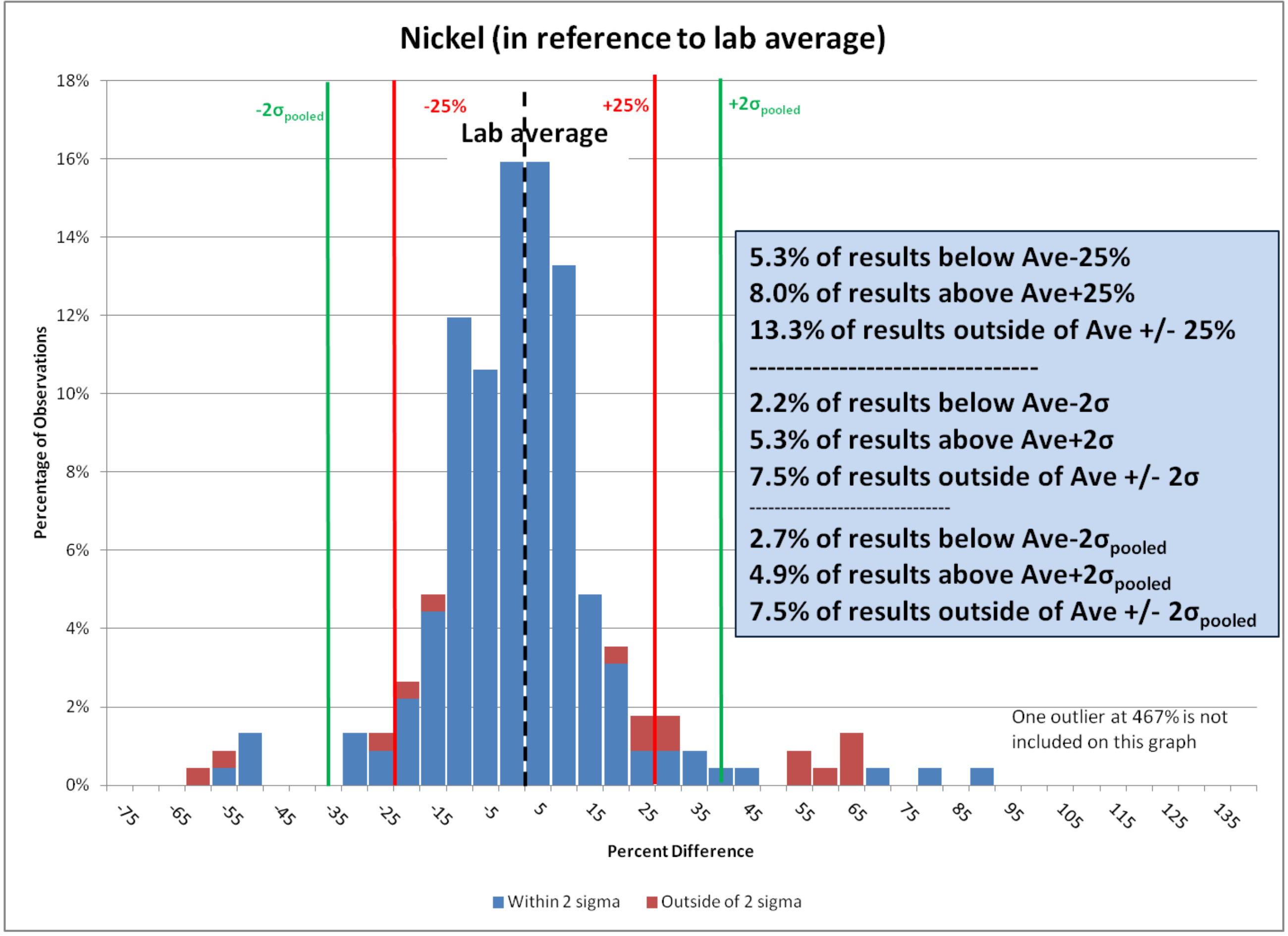
One outlier at 344% is not included on this graph



Nickel (in reference to T)

7.1% of results below T-25%
7.1% of results above T+25%
14.2% of results outside of T +/- 25%
--------------------------------
5.8% of results below T-2σ
4.9% of results above T+2σ
10.6% of results outside of T +/- 2σ
--------------------------------
3.1% of results below T-2$\sigma_{pooled}$
4.9% of results above T+2$\sigma_{pooled}$
8.0% of results outside of T +/- 2$\sigma_{pooled}$

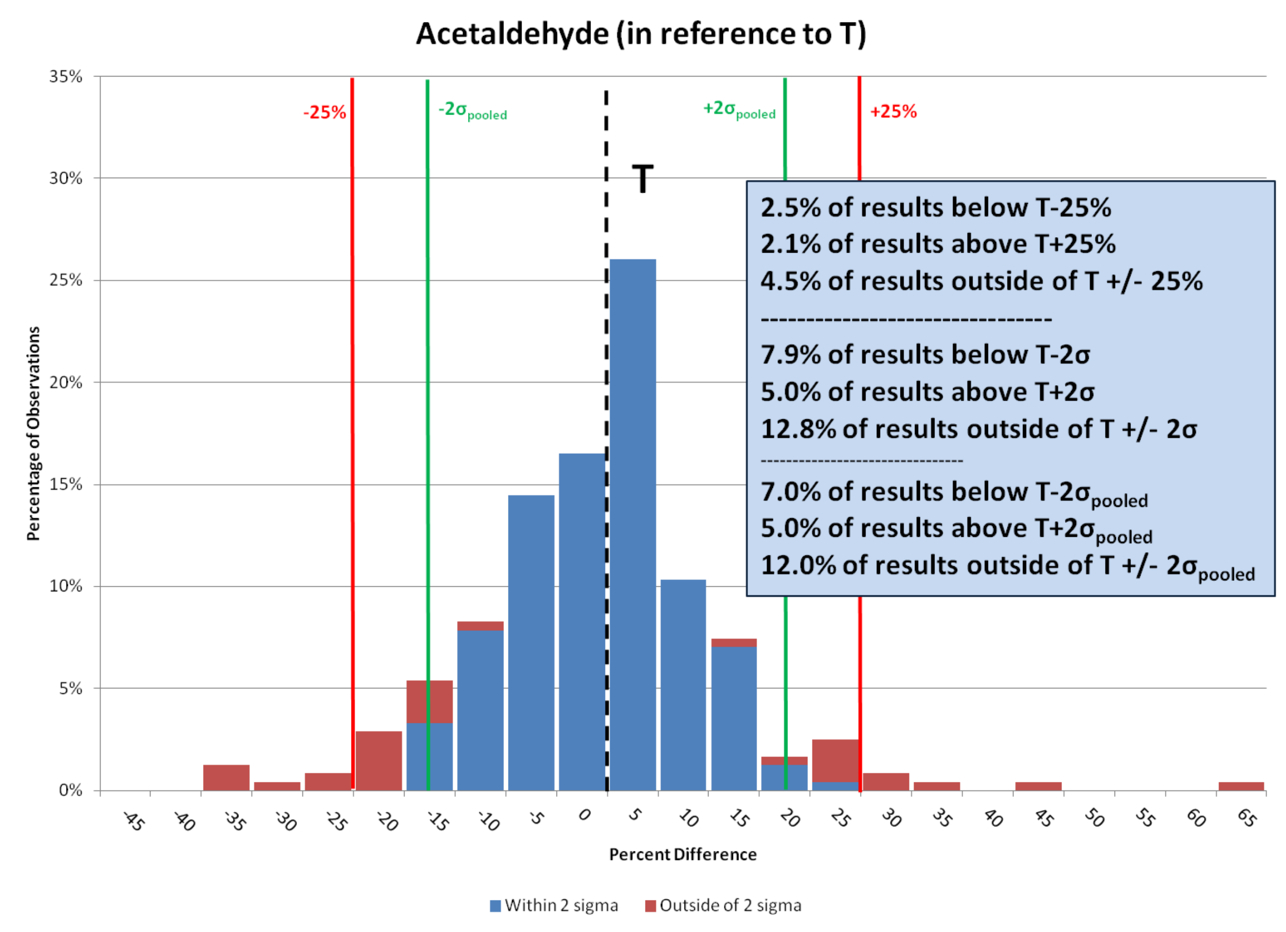
One outlier at 429% is not included on this graph



Nickel (in reference to lab average)

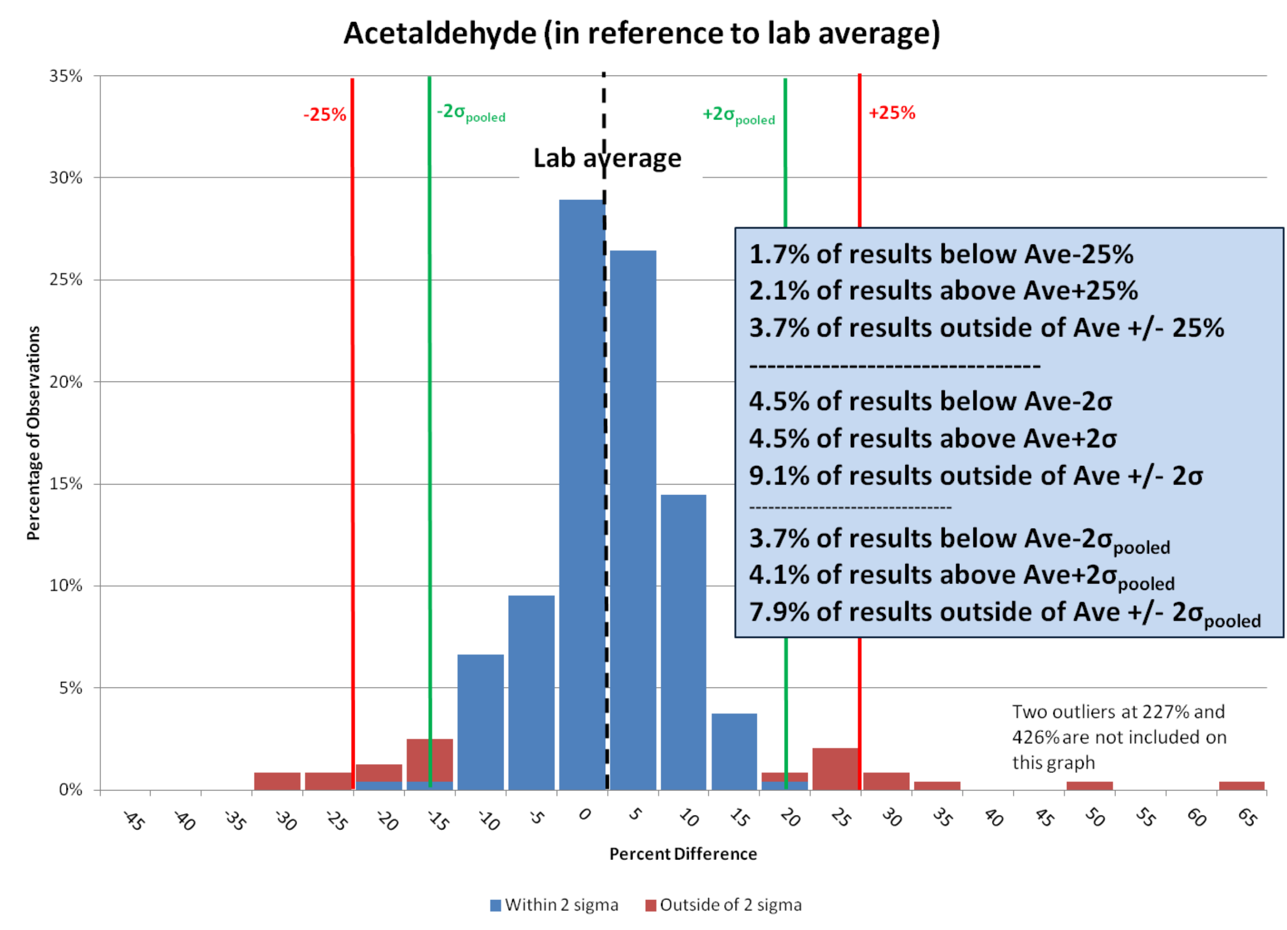
5.3% of results below Ave-25%
8.0% of results above Ave+25%
13.3% of results outside of Ave +/- 25%
--------------------------------
2.2% of results below Ave-2σ
5.3% of results above Ave+2σ
7.5% of results outside of Ave +/- 2σ
--------------------------------
2.7% of results below Ave-2$\sigma_{pooled}$
4.9% of results above Ave+2$\sigma_{pooled}$
7.5% of results outside of Ave +/- 2$\sigma_{pooled}$

One outlier at 467% is not included on this graph



Acetaldehyde (in reference to T)

2.5% of results below T-25%
2.1% of results above T+25%
4.5% of results outside of T +/- 25%
--------------------------------
7.9% of results below T-2σ
5.0% of results above T+2σ
12.8% of results outside of T +/- 2σ
--------------------------------
7.0% of results below $T-2\sigma_{pooled}$
5.0% of results above $T+2\sigma_{pooled}$
12.0% of results outside of T +/- $2\sigma_{pooled}$



Acetaldehyde (in reference to lab average)

1.7% of results below Ave-25%
2.1% of results above Ave+25%
3.7% of results outside of Ave +/- 25%
--------------------------------
4.5% of results below Ave-2σ
4.5% of results above Ave+2σ
9.1% of results outside of Ave +/- 2σ
--------------------------------
3.7% of results below $Ave-2\sigma_{pooled}$
4.1% of results above $Ave+2\sigma_{pooled}$
7.9% of results outside of Ave +/- $2\sigma_{pooled}$

Two outliers at 227% and 426% are not included on this graph

# APPENDIX B

## SCATTERPLOTS OF THE STANDARD DEVIATION OF PERCENT DIFFERENCE ($PD_2$) VERSUS YEAR/QUARTER FOR REMAINING FOUR METALS

(Plots for arsenic and nickel were found in Figures 2 and 4.)

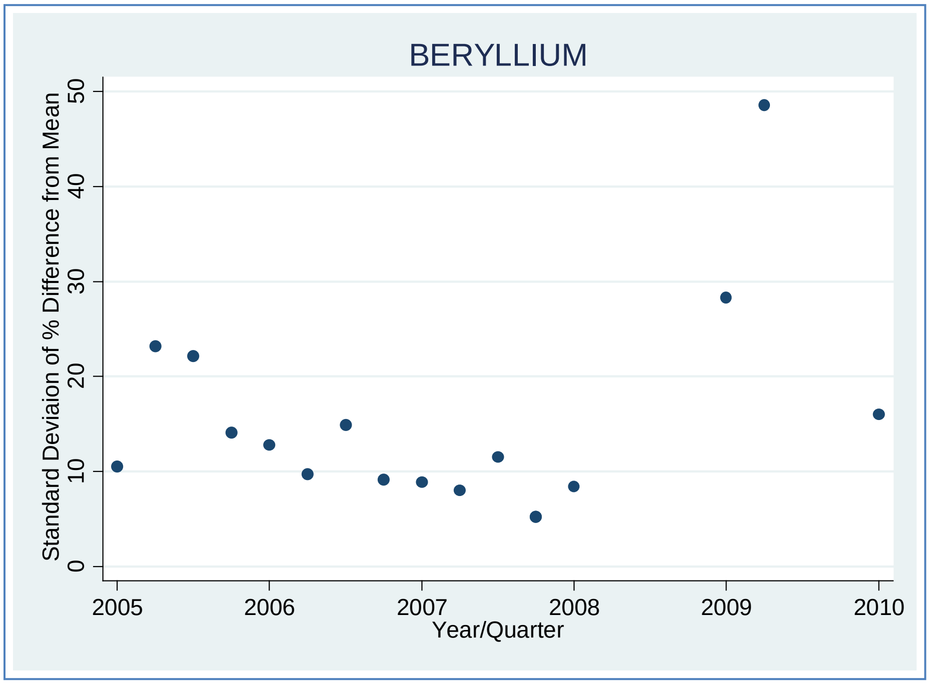


**Figure B1. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Beryllium**

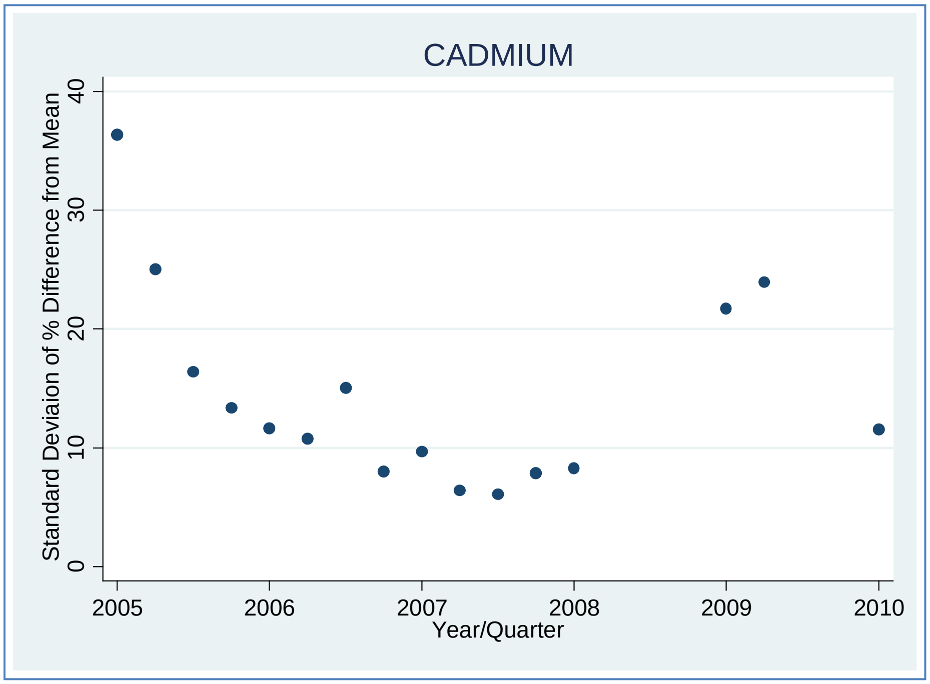


**Figure B2. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Cadmium**

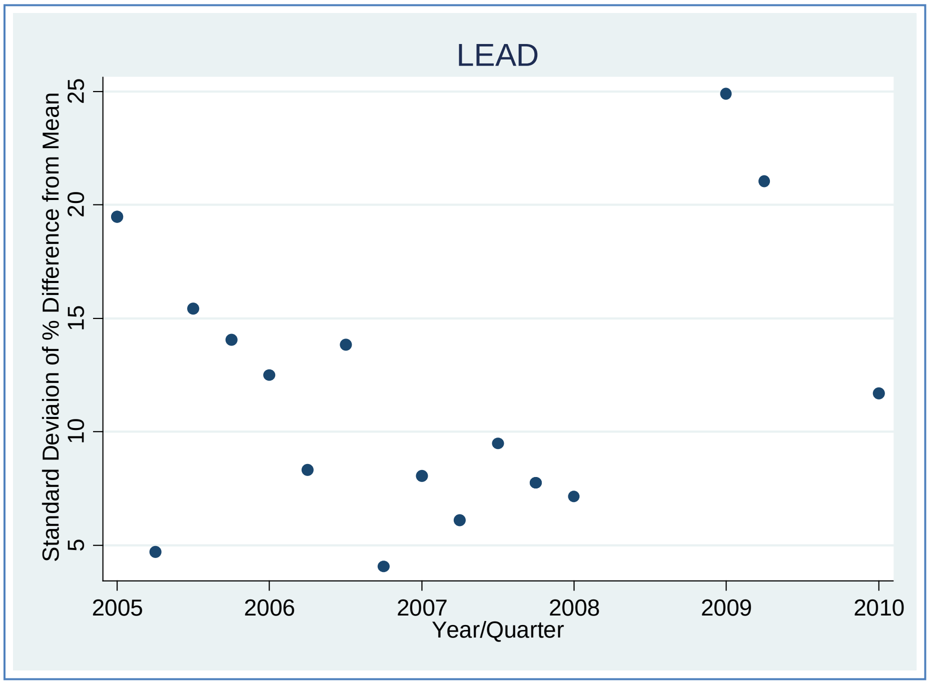


**Figure B3. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Lead**

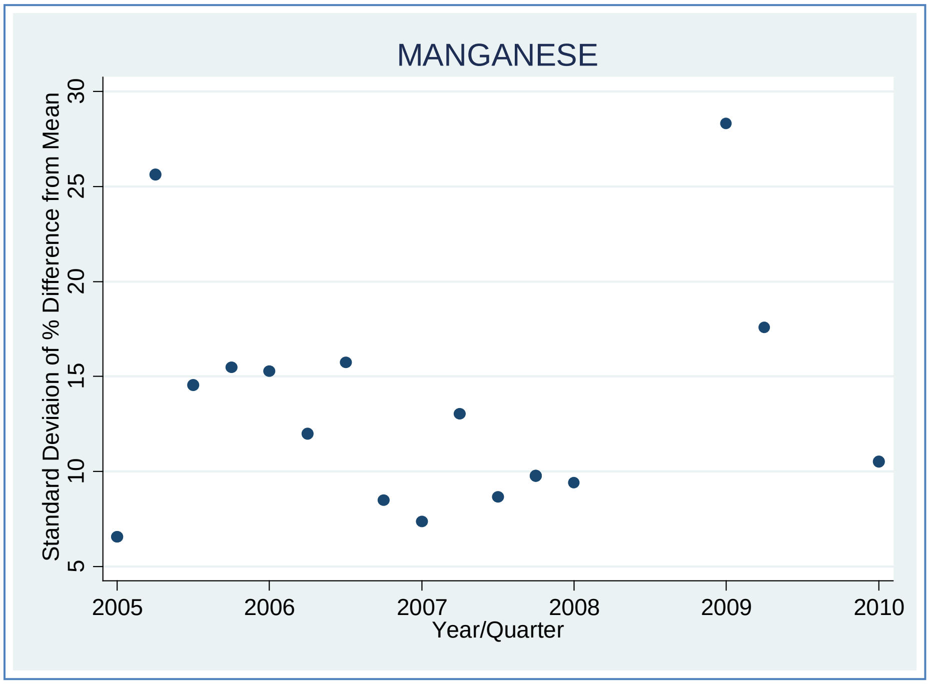


**Figure B4. Plot of the Standard Deviation of $PD_2$ Versus Year/Quarter for Manganese**

# APPENDIX C

## SCATTERPLOTS OF PROFICIENCY TEST SAMPLE VALUES VERSUS BATCH-SPECIFIC LAB AVERAGE FOR SIX ANALYTES CONSIDERED IN PHASE II AND THREE CONDITIONS FOR DATA ACCEPTANCE

Definition of Condition:

- A: The outliers identified in Phase I by Grubb's test performed at the 1% significance level were excluded.
- B: In addition to A,
  - the measurements for the 10 lab/analyte classes identified in Table 3 (Section 2.2.1) were excluded,
  - acrolein data from 2006 were excluded (due to a considerably high degree of variability compared to subsequent years).

For the two metals (arsenic and nickel), $\sigma_{\text{pooled}}$ was estimated under an additional condition:

- C: In addition to B, all data from 2009 were excluded

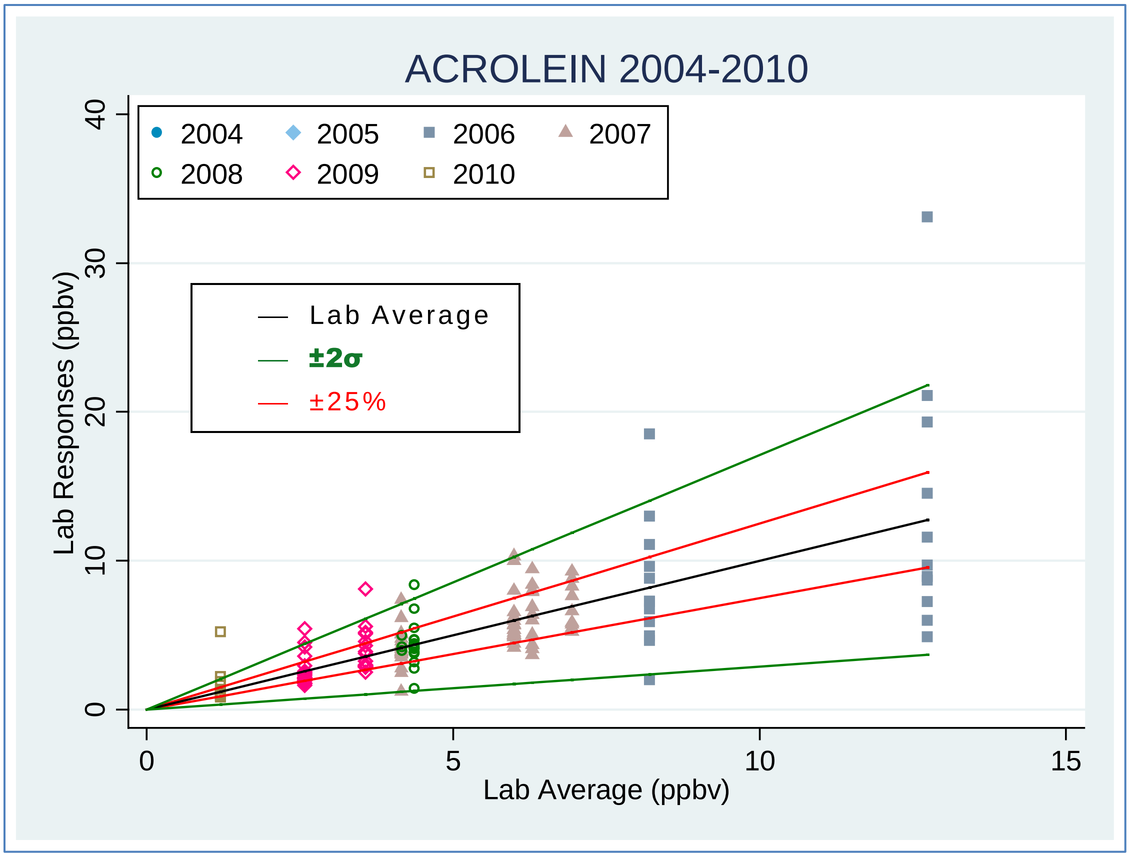


Figure C1. Plot of Lab Response Versus Lab Average for Acrolein, Condition A

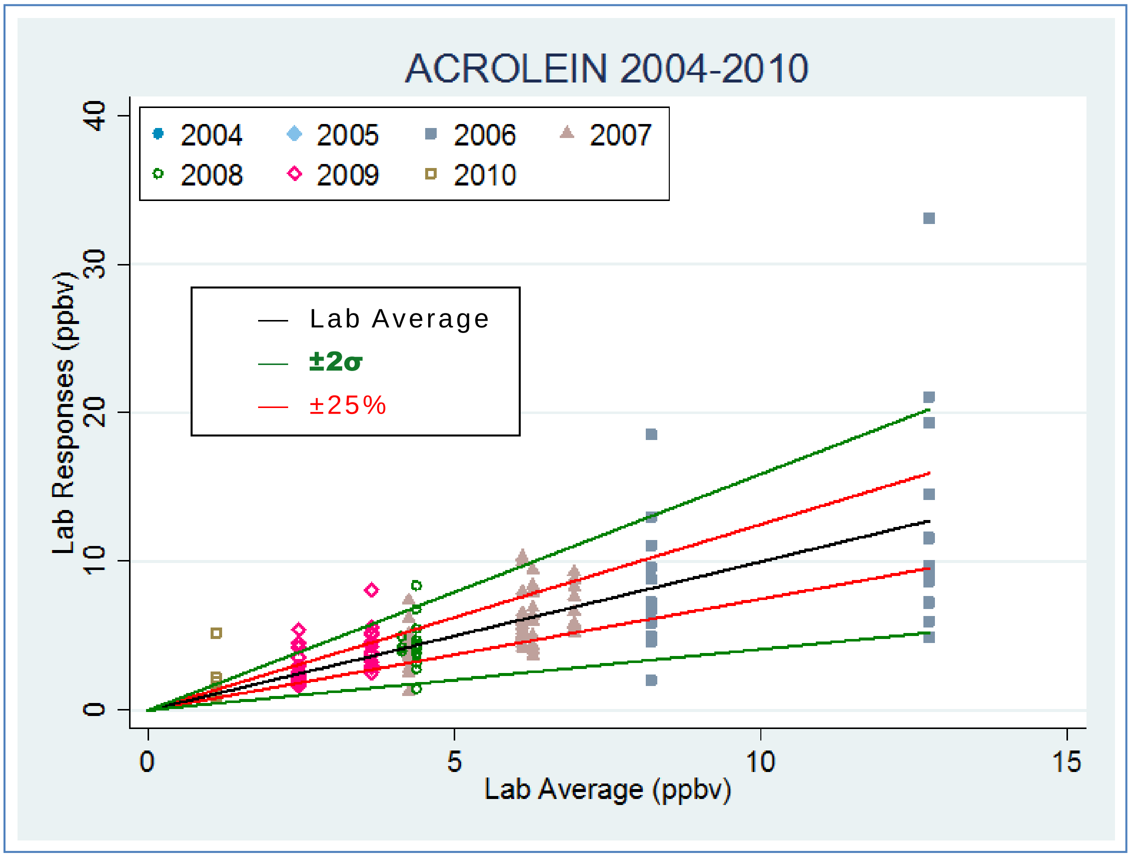
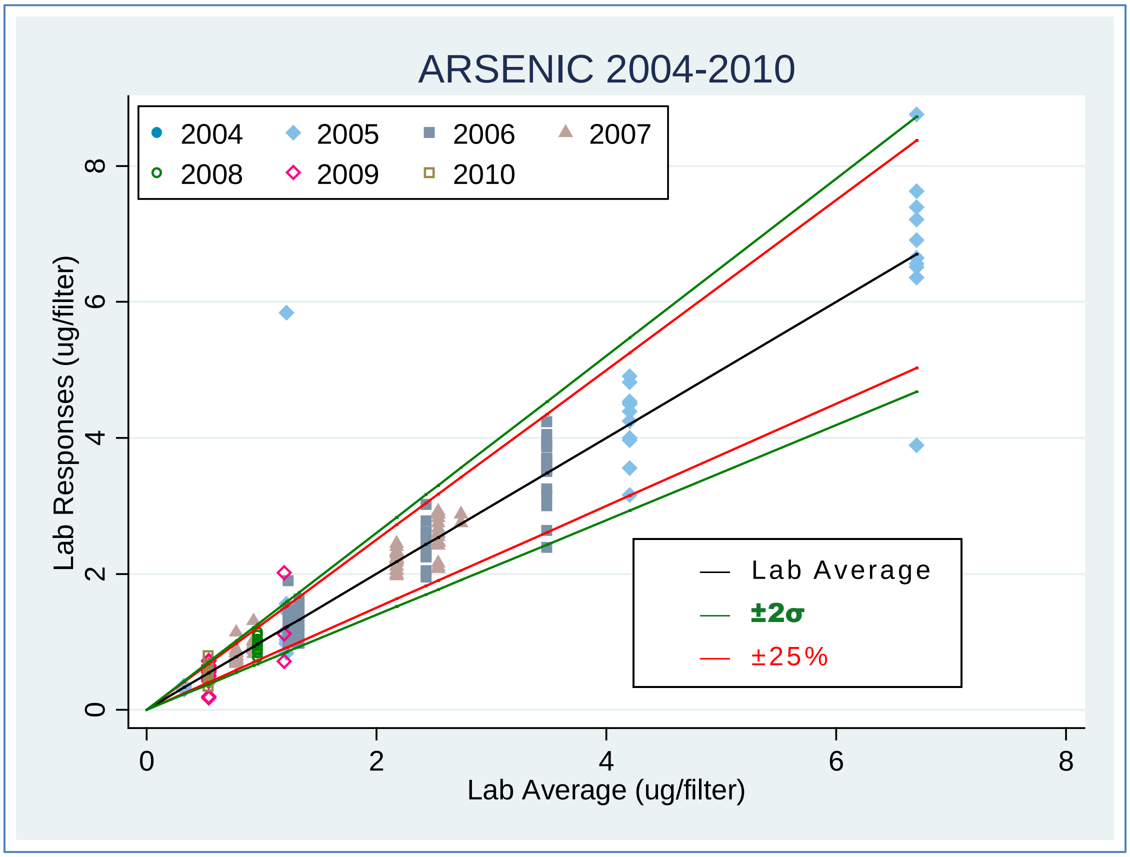
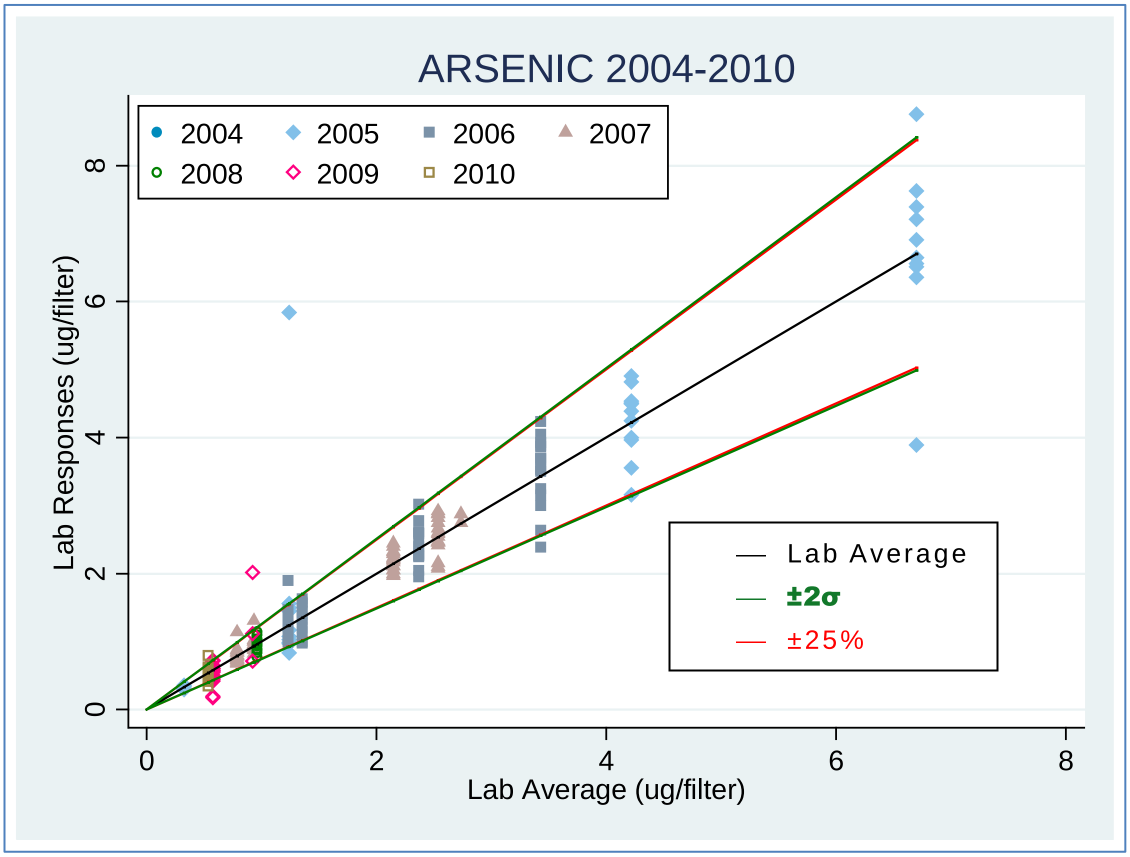


Figure C2. Plot of Lab Response Versus Lab Average for Acrolein, Condition B



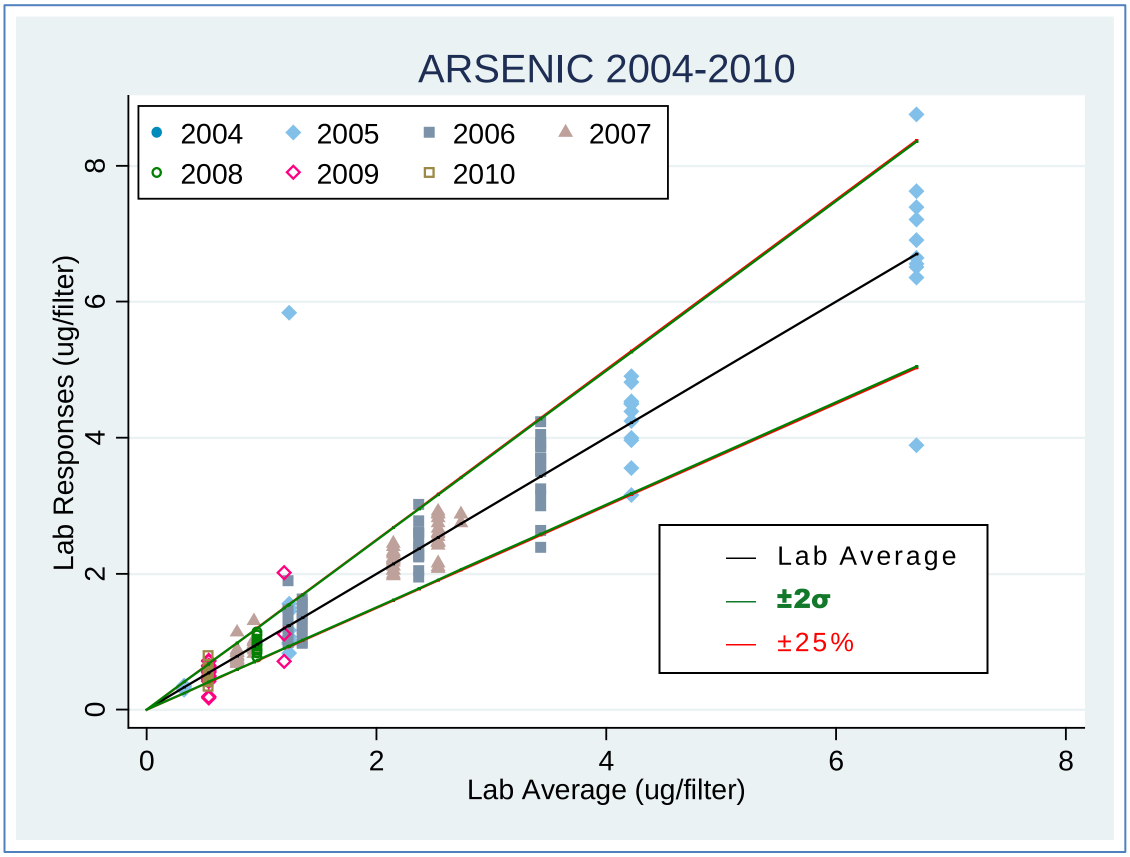
One outlier (Lab Response=20, True=5.88, Year=2005) is not included in the graph

**Figure C3. Plot of Lab Response Versus Lab Average for Arsenic, Condition A**



One outlier (Lab Response=20, True=5.88, Year=2005) is not included in the graph

**Figure C4. Plot of Lab Response Versus Lab Average for Arsenic, Condition B**



One outlier (Lab Response=20, True=5.88, Year=2005) is not included in the graph

**Figure C5. Plot of Lab Response Versus Lab Average for Arsenic, Condition C**

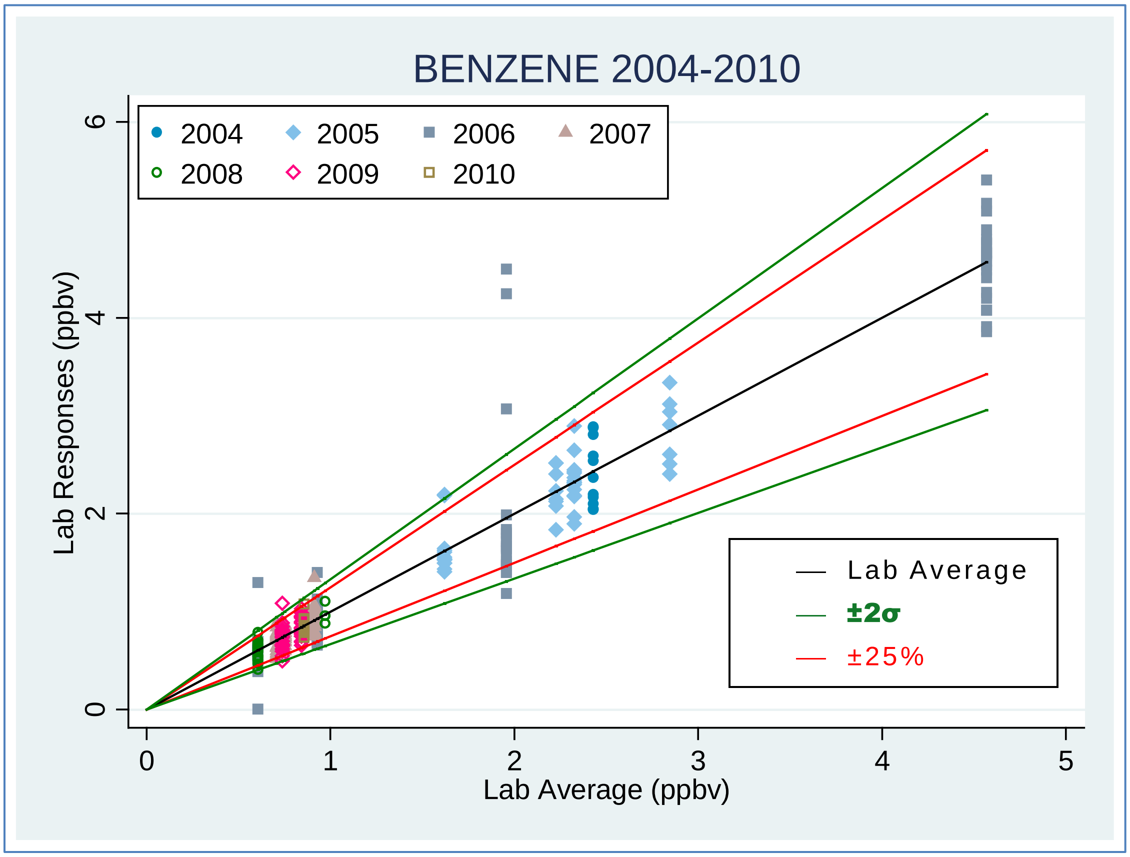


Figure C6. Plot of Lab Response Versus Lab Average for Benzene, Condition A

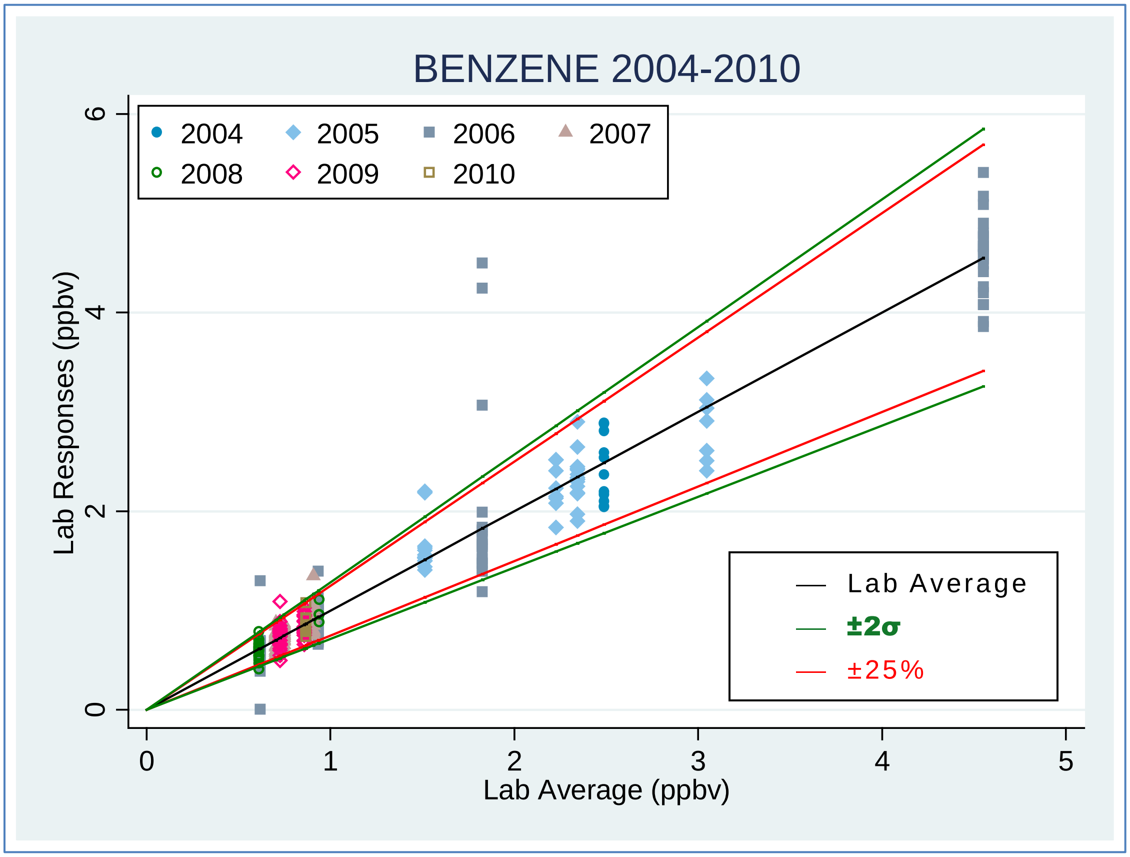


Figure C7. Plot of Lab Response Versus Lab Average for Benzene, Condition B

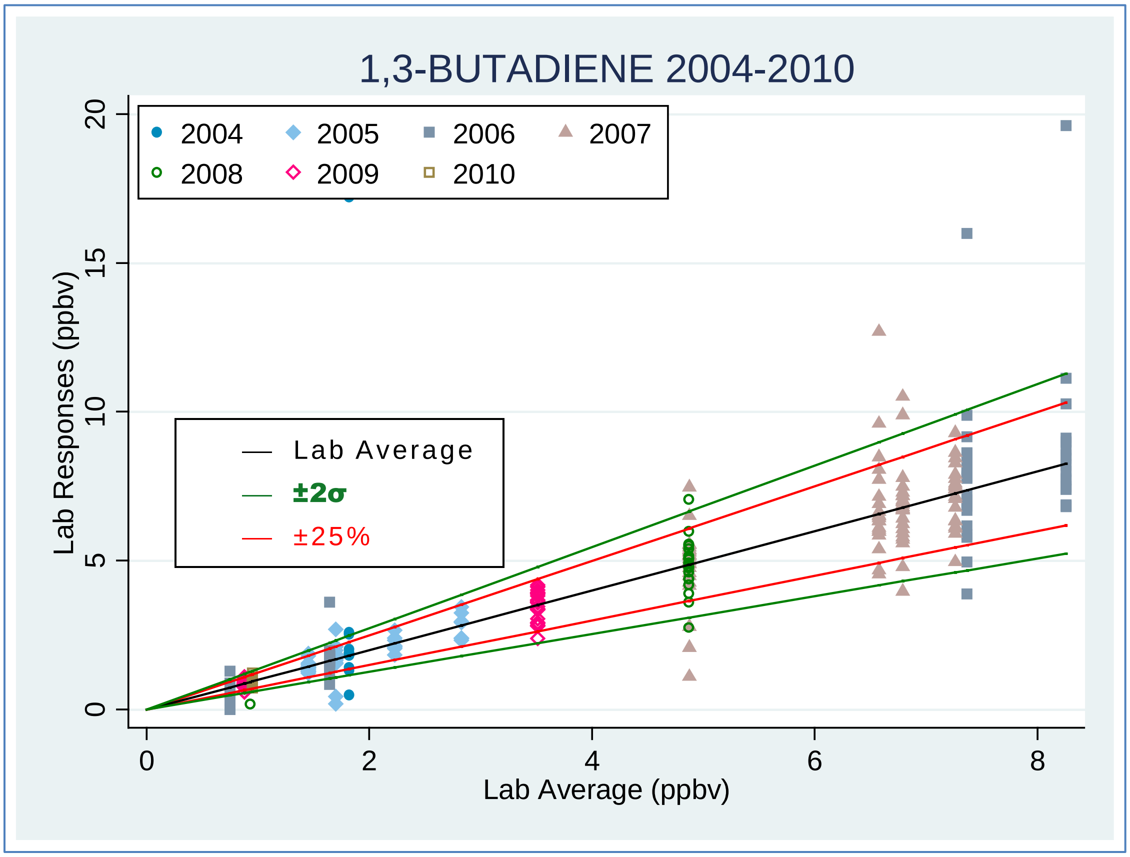


**Figure C8. Plot of Lab Response Versus Lab Average for 1,3-Butadiene, Condition A**

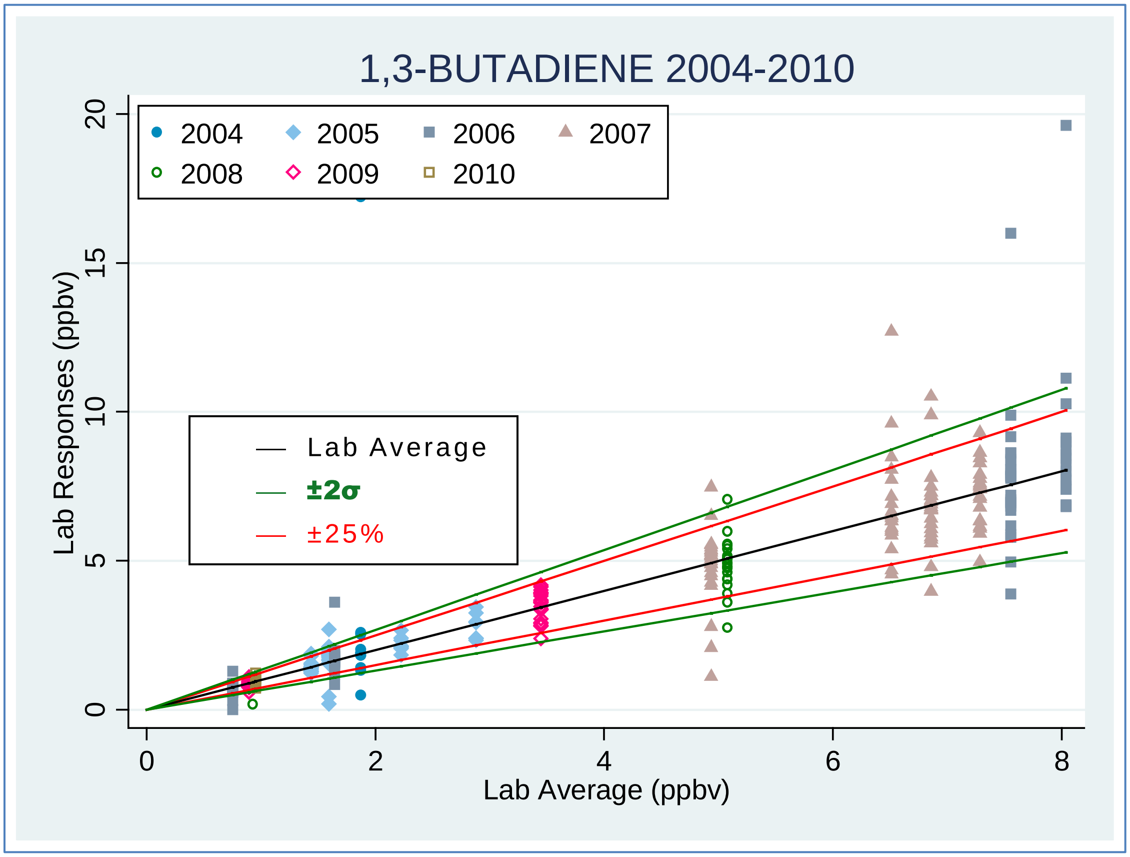


**Figure C9. Plot of Lab Response Versus Lab Average for 1,3-Butadiene, Condition B**

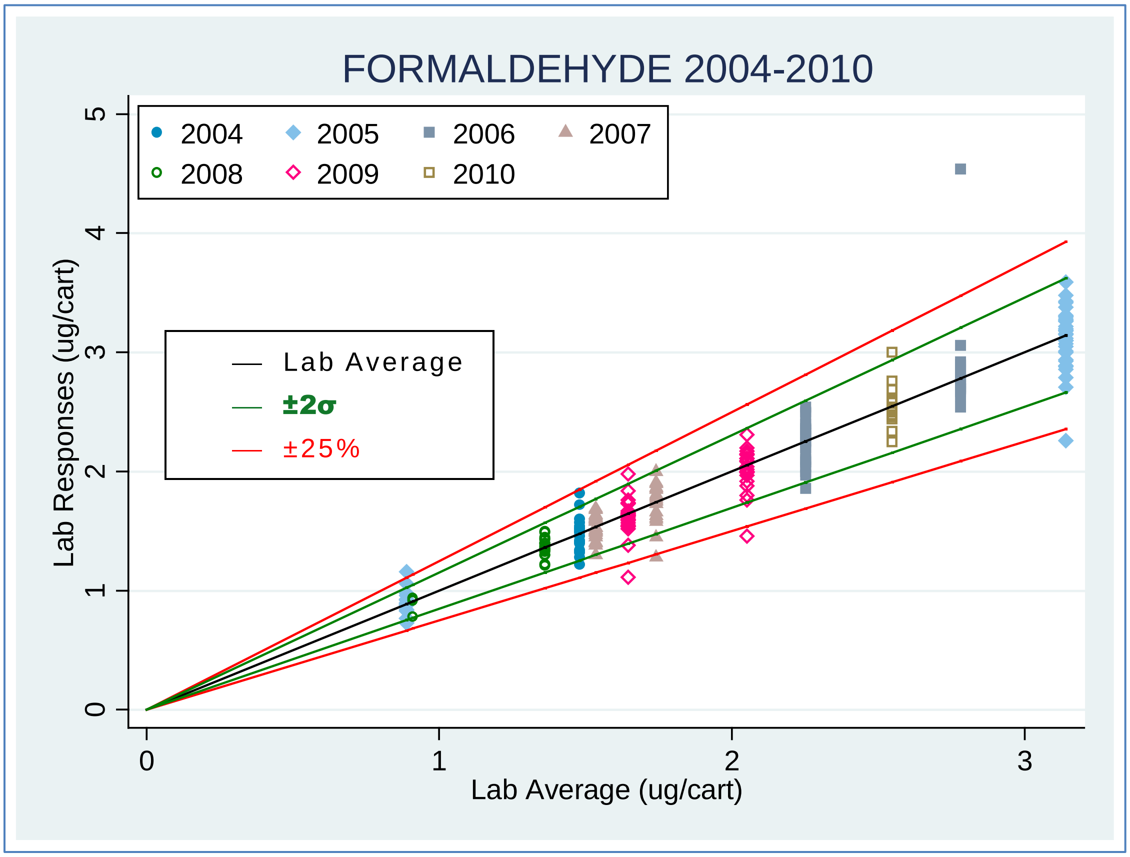


Figure C10. Plot of Lab Response Versus Lab Average for Formaldehyde, Condition A

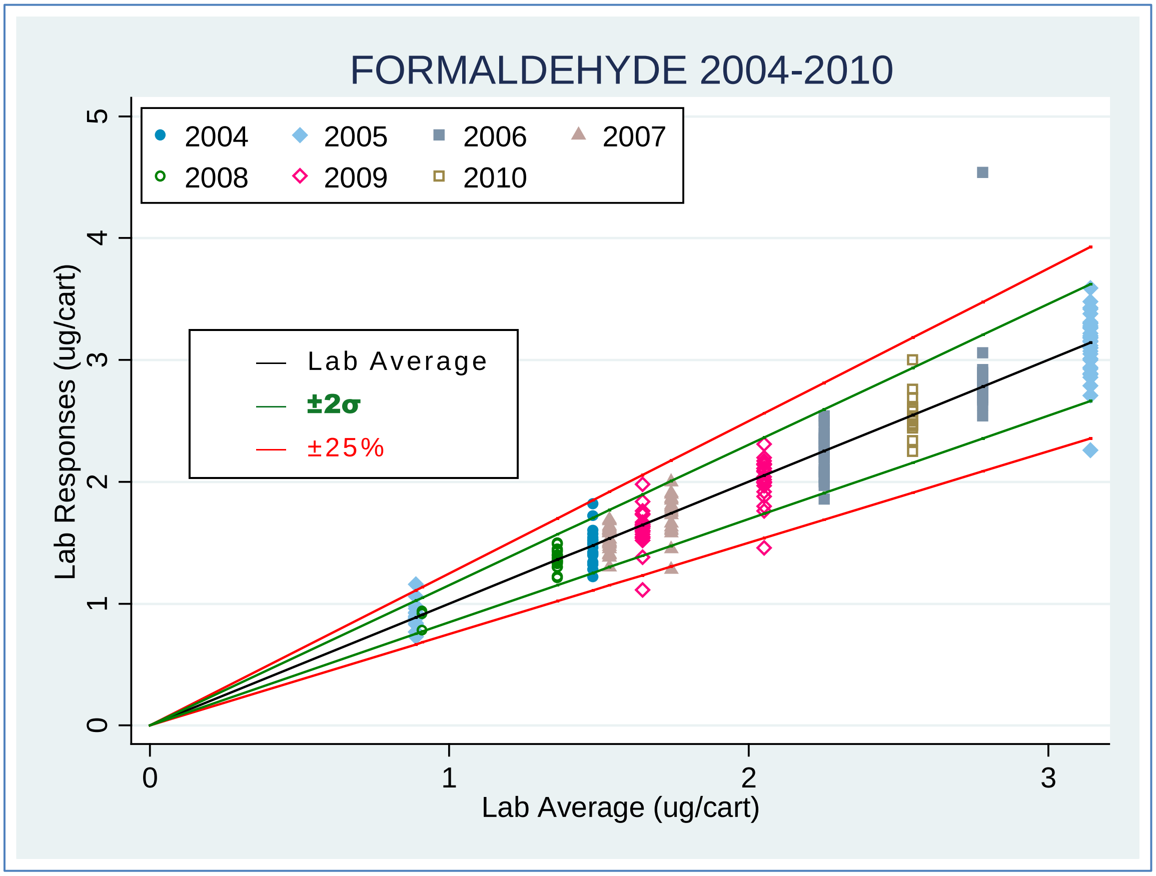


Figure C11. Plot of Lab Response Versus Lab Average for Formaldehyde, Condition B

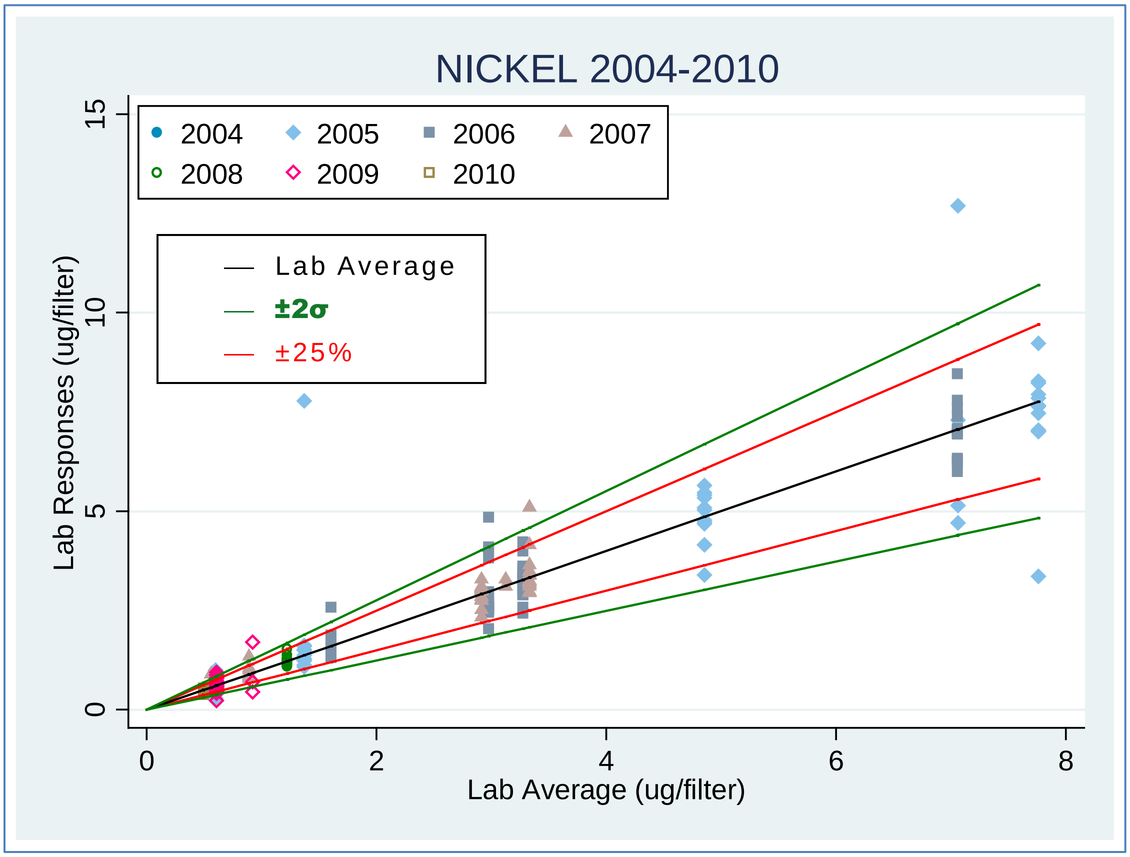


**Figure C12. Plot of Lab Response Versus Lab Average for Nickel, Condition A**

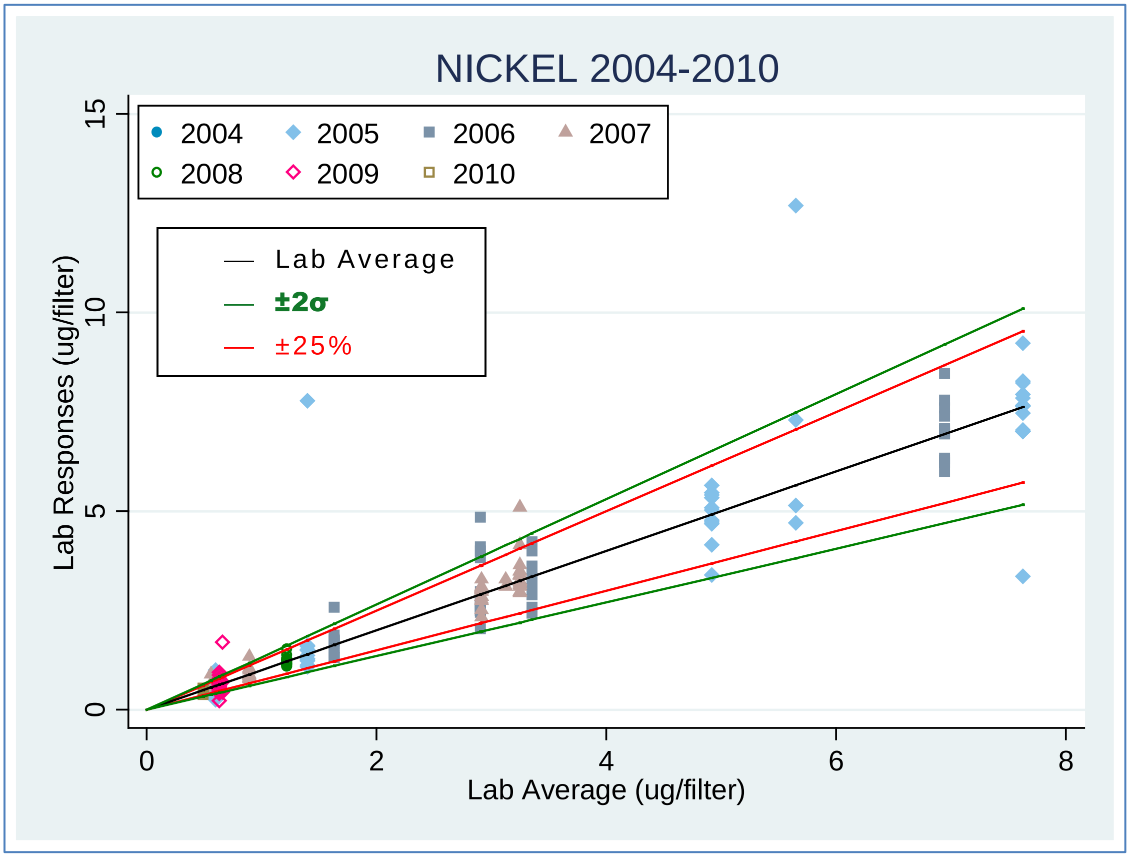


**Figure C13. Plot of Lab Response Versus Lab Average for Nickel, Condition B**

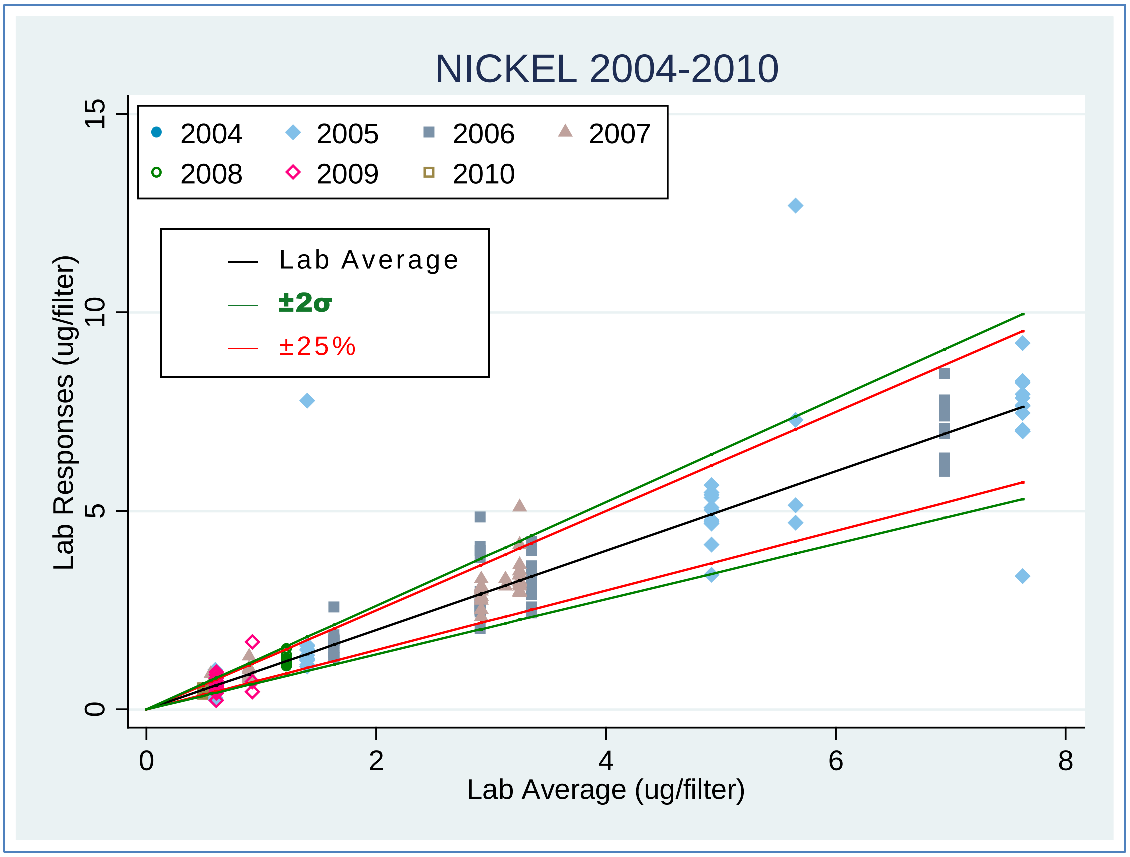


**Figure C14. Plot of Lab Response Versus Lab Average for Nickel, Condition C**