Notes on NH3 methods comparison/instrument change

Elgin Perry
eperry@chesapeake.net
410-535-2949
4/10/2012

In a previous note to Rick Hoffman (see the end of this document), I reported that I that I saw no major issues with the Ammonia Method/Instrument Change Study and results. After conducting my own analysis of these data, I find that there are issues including:

1. dependence among observations,
2. difference between methods changes with the mean, and
3. significance of difference is overstated by original analysis.

On the whole, I do not feel that these issues invalidate the conclusion that the Lachat method yields slightly higher measurements of NH3 than does the Skalar method. However, issue 2 suggests that if adjustment factors are developed for purposed of comparing across a change in methods, then these factors need to take mean NH3 into account, and issues 1 and 3 suggest that there are random factors causing differences among observations that are not well characterized by this study. In the future, it would be prudent to implement a study design that more completely characterizes these random factors.

In what follows, I present analyses to illustrate these three issues and discuss consequences for this and future method comparison studies.

Issue 1. Dependence Among Observations.

The first issue of dependence among observations becomes apparent in a run sequence plot of the differences between methods (Lachat - Skalar) (Figure 1). It is clear that there are sequences of observations that have a strong tendency to deviate to either the positive side or negative side of the mean difference. This tendency tends to co-occur with the dates of the chemical analyses. The observations for comparisons run on Dec. 3 (Lachat date) and Dec. 2 (Skalar date) are shown is red as an illustration of this bias. The fact that observations made together on a day tend to be more similar to each other than observations made on different days suggests that there is a random factor that affects all observations on one day similarly and that the effect of this random factor changes when the experiment is set up on a different day.

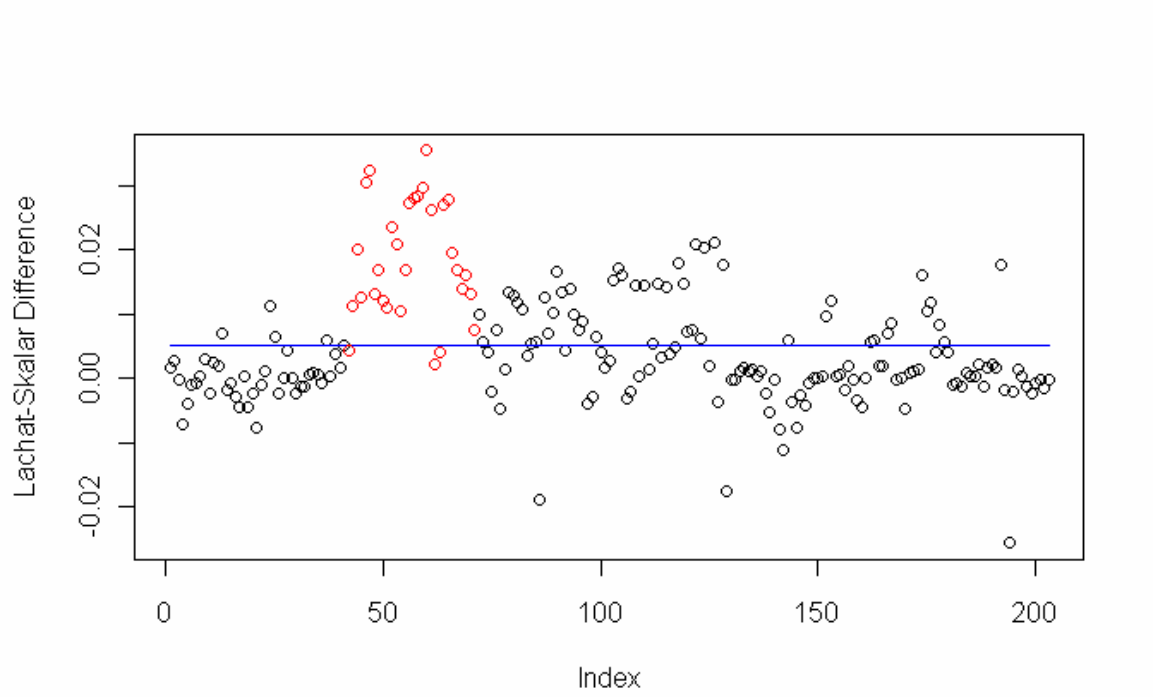


Figure 1. Run-sequence plot of differences between Lachat and Skalar (Lachat - Skalar) . The observations for the Dec. 3 to Dec. 2 comparison are in red.

Based on this observation of dependence, I created a variable called 'run' that takes a unique value for each unique combination of Lachat date and Skalar date (Table 1). In turn, the difference data are presented in box plots by run (Figure 2).

Table 1. Run dates defined by unique combinations of Lachat analysis date and Skalar analysis date.

run	Lachat date	Skalar date
1	12/19/2008	12/11/2008
2	12/3/2008	12/2/2008
3	5/8/2008	5/6/2008
4	5/15/2008	5/23/2008
5	5/15/2008	6/4/2008
6	12/4/2009	12/10/2009
7	2/10/2010	2/4/2010
8	2/10/2010	1/25/2010
9	7/13/2010	7/13/2010

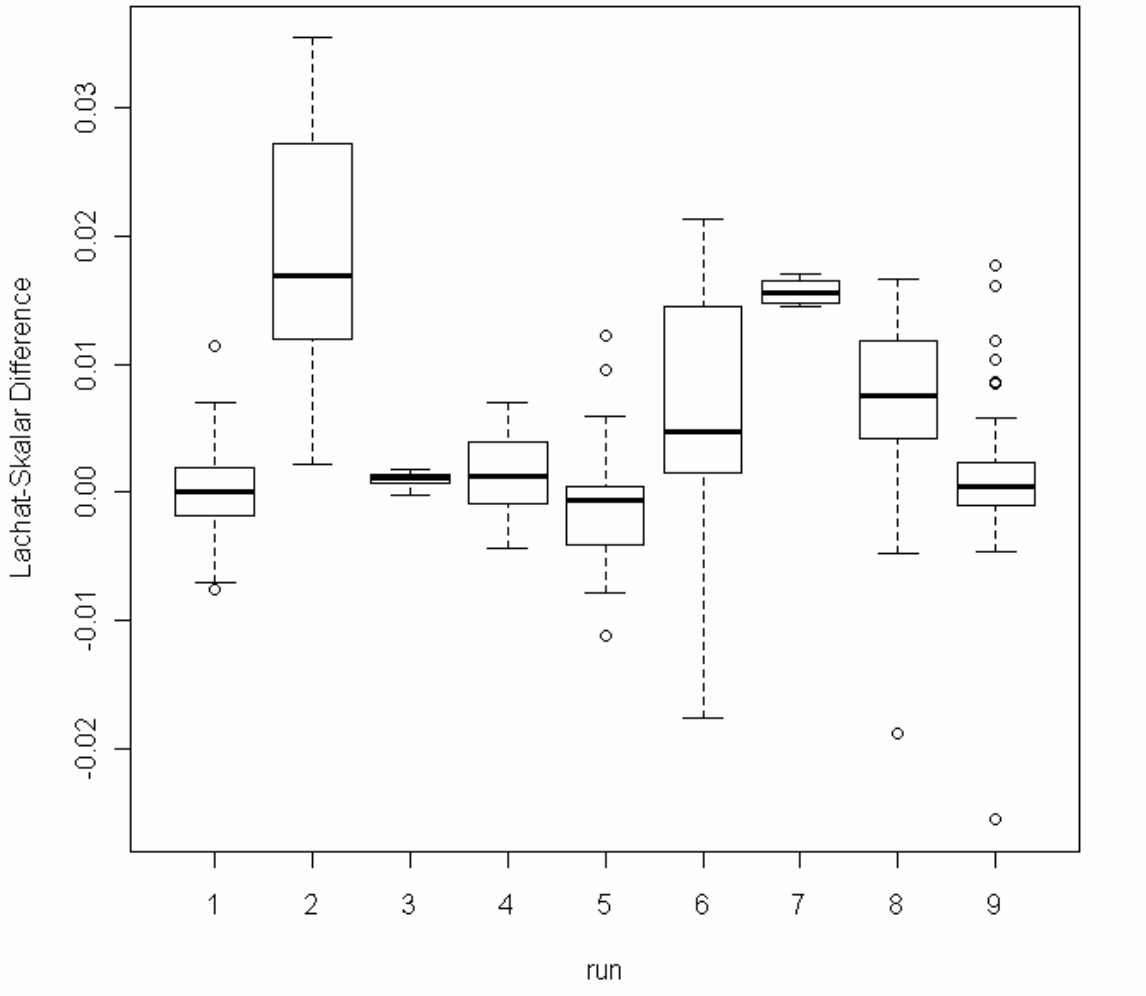


Figure 2. Box plots of the Lachat-Skalar difference as an function of analysis run dates (Table 1.)

There are some runs where the differences seems centered on zero, e.g. runs 1 and 9, and others where the difference seems positive, e.g. runs 2 and 7. There is only one date (run 5) where the body of the differences tends toward negative. That boxes do not overlap is an indication of a date to date random factor.

Issue 2. Difference Between Methods Changes with the Mean.

The differences between the Lachat and Skalar methods tend to be a function of the mean NH3 concentration (Figure 3). In the interval from (0, 0.03 mg/l) the difference between the methods does not appear to be significantly different from zero and thus at low concentration it is acceptable to consider the methods equivalent. However, at about 0.03 mg/l, the difference between methods begins to increase rapidly to where the Lachat method measure NH3 is about 0.01 mg/l greater than the Skalar measurement. Based on this, if one were to make an adjustment to a time series that includes data from both instruments, this adjustment should be a function of the observed NH3.

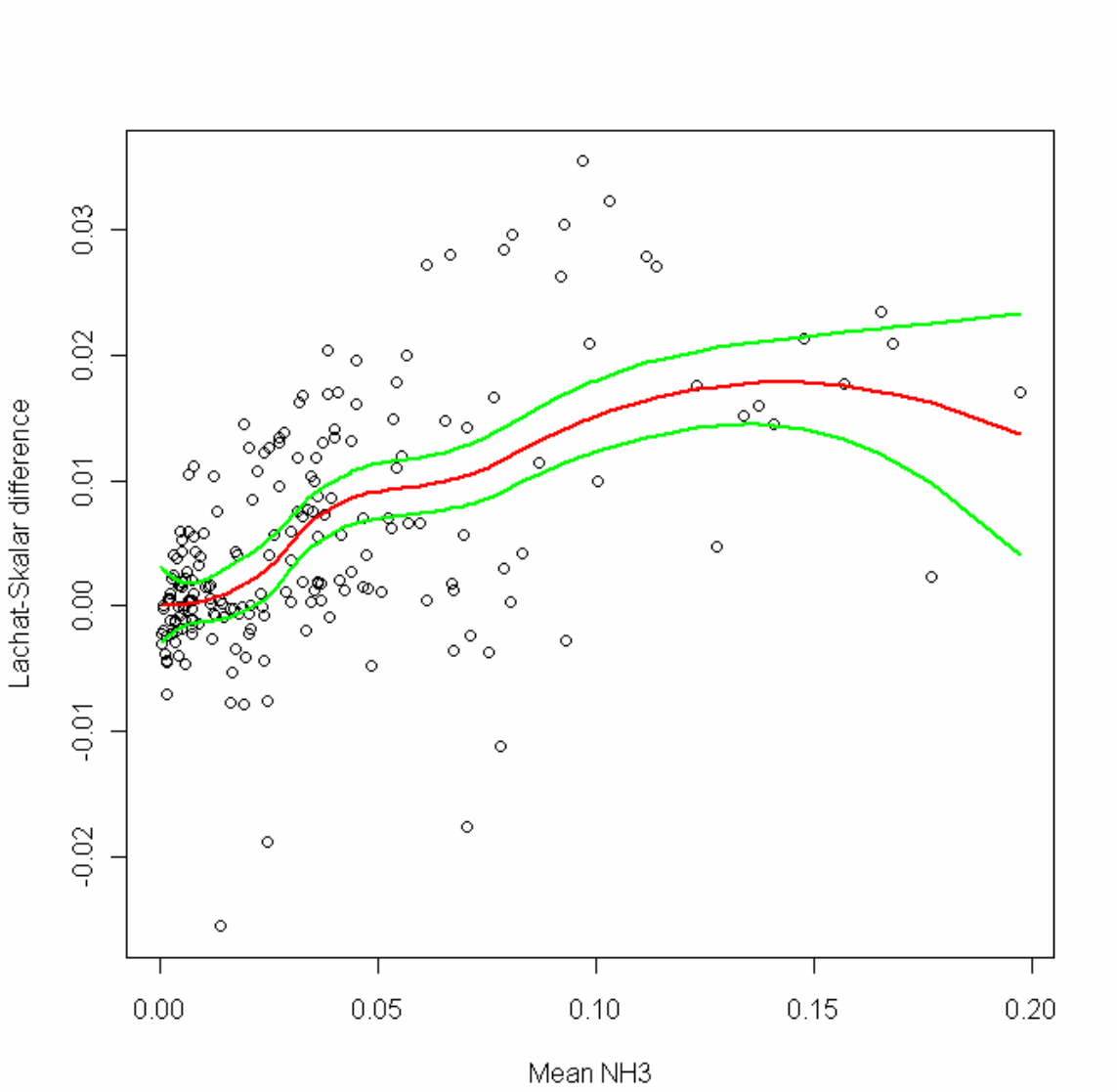


Figure 3. The Lachat-Skalar differences plotted as a function of mean (Lachat,Skalar) NH3 with a loess regression line (red) and confidence intervals (green).

Looking at the difference of logarithms plotted against mean concentration (Figure 4), it appears that the log difference is nearly constant over the range of observations which suggests that the difference between instruments might be computed as a simple proportion. The mean of the Log-differences is 0.1402189. Note that observations when the Lachat method produced negative values were excluded from this calculation. This result suggests that the Lachat observations should be multiplied by 0.8691679 in order to adjust them to match the Skalar observations (Figure 5).

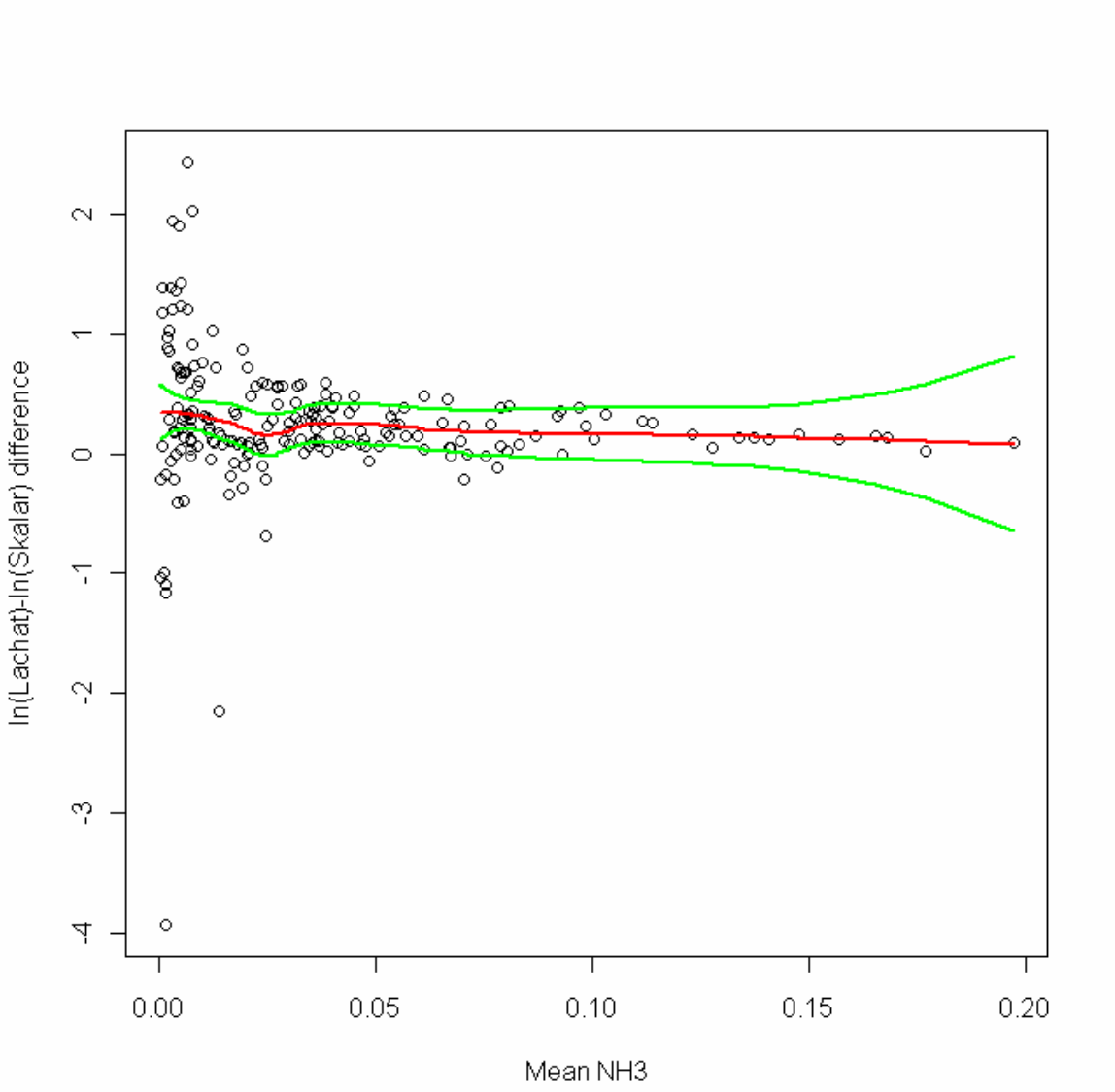


Figure 4. Differences in logarithm transformed measurements from Lachat and Skalar methods plotted against mean concentration.

To assess the efficacy of this adjustment, it was applied to the Lachat data for this study and compare to the unadjusted Lachat data (Figure 5). The adjusted

data does appear to be fairly well centered on the one-to-one line. However, the sample size at high concentrations is sparse and may not provide a good assessment of this adjustment.

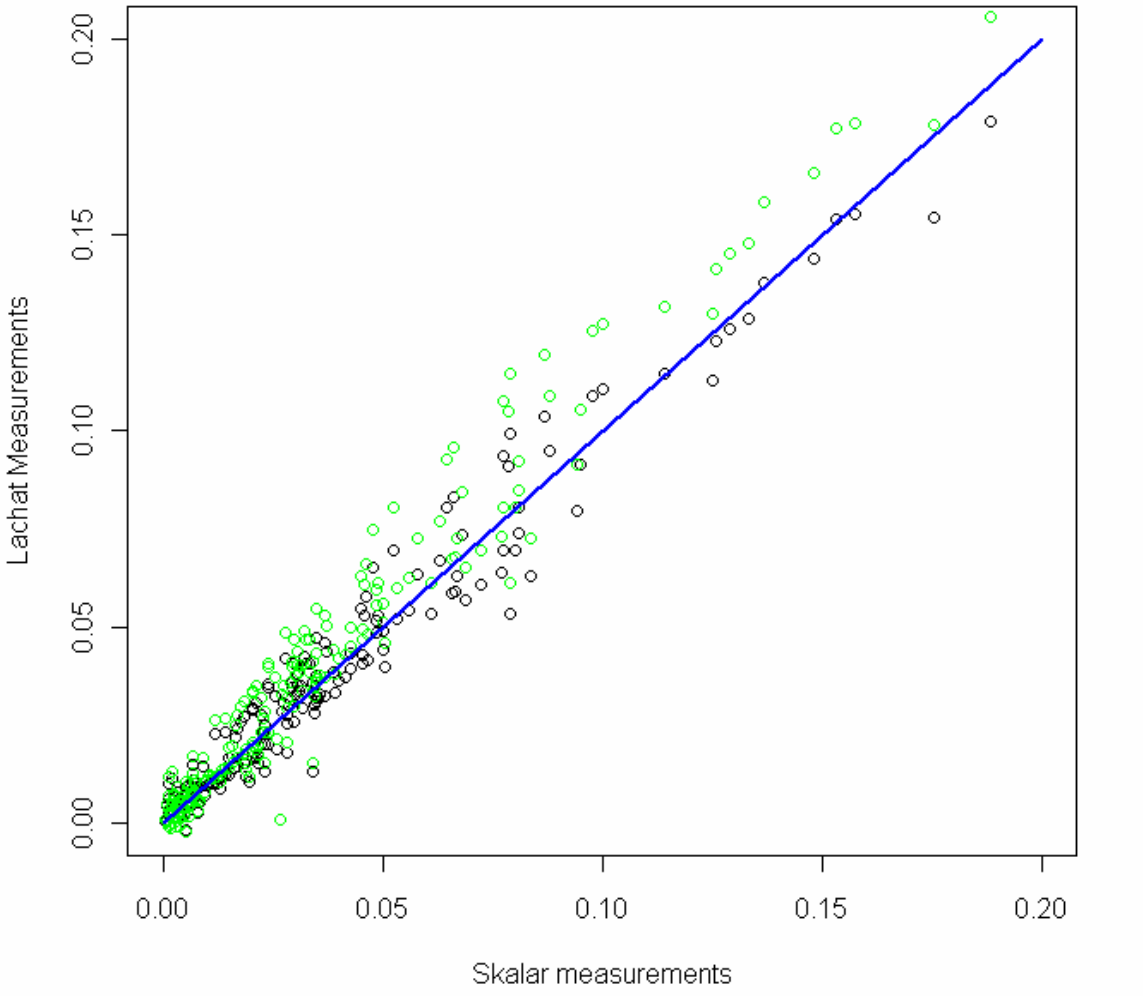


Figure 5. Skalar measurements plotted versus Lachat measurements (green circle) and adjusted Lachat measurements (black circle). The one-to-one line is shown in blue.

Issue 3. Significance of Difference is Overstated by Original Analysis.

This issue of overstating the statistical significance of the difference is related to issue 1, the dependence of method differences that were taken together in a run defined by pairs of Lachat-Skalar dates. It appears that there is some random factor that differs from run to run and that observations taken within a run are pseudo-replicates or repeated measures of this random run to run factor. Because of the lack of independence among observations within a run, the assumption of independence required for the paired t-test analysis is not met. A more appropriate analysis is to perform a nested analysis of variance where run is the highest level factor and sample pair is nested within run. In this analysis, it is the runs that are assumed to be independent and the samples nested within runs are assumed dependent with respect to the among runs factor. It is the number of runs that determines sample size (9) rather than the number of samples (203). This reduction of sample size is the biggest factor affecting the significance of the test. The results from this nested analysis are shown below. The estimate of the difference between methods is 0.005488 which is quite similar to the difference of 0.005229 reported in *Technical Procedure 2-600.pdf* . The p-value for this difference reported here is 0.0458 which is just below the 0.05 cut off for significance but much larger than the p-value reported from the paired t-test. This indicates much more of a border-line case than would be interpreted from the <0.0001 p-value reported from the paired t-test.

Results from mixed model analysis using R

```
Linear mixed model fit by REML
Formula: nh3diff ~ 1 + (1 | frun)
   Data: mc1
   AIC   BIC logLik deviance REMLdev
 -1407 -1397  706.5    -1423   -1413
Random effects:
 Groups   Name        Variance   Std.Dev.
 frun     (Intercept) 4.5210e-05 0.0067238
 Residual             4.6568e-05 0.0068241
Number of obs: 203, groups: frun, 9

Fixed effects:
            Estimate Std. Error t value   p-value
(Intercept) 0.005488   0.002323   2.362    0.0458
```

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna,
Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Improvements for Experimental Design

As is often the case, once the data are in hand it is easy to see how the experimental design might have been improved. It is instructive to review these potential improvements as a means of improving future experiments to assess methods changes.

First we note the importance of the variance among runs which is much larger than the variance among samples within runs. The larger variance of the among runs factor coupled with the dependence of samples within run implies that it is more important to assess as many runs as possible.

Second we note that the magnitude of the difference is a function of sample concentration. In this experiment we have few observations at high concentration and a large number a low concentration. It would be an improvement to have nearly equal numbers of samples throughout the expected concentration range. This could be achieved by spiking samples if this variation is difficult to obtain naturally.

Thirdly, an experiment based on analysis of paired samples will disclose if two methods are producing different results, but will not inform us on which of the two methods is more correct. To address this issue, samples based on dilution of standards could be prepared and analyzed in pairs by the two methods. If it is desired to perform this test in the presence of natural interfering constituents, a natural sample could assayed before and after the addition of a known quantity and the methods can be compared on which more accurately measured the increase.

Analysis of data = Second nh3 study.xls

These data yield results similar to the above.

There are no dates reported with the second study and thus the issue of dependence within dates cannot be addressed. I did a runs test for dependence and find that these data as stored in the spread sheet show autocorrelation which indicates dependence.

```
runs.test(mc2$nh3diff, alternative = "positive.correlated")

        Runs Test - Positive Correlated

data:  mc2$nh3diff
Standardized Runs Statistic = -1.9234, p-value = 0.02722
```

Because dates are not reported with these data, they are not as useful for assessing the methods change. The assessments that can be done without the date information is presented below.

These data also show dependence of the difference on the mean (Figure 6). As the mean increases the difference tends to increase as well. In these data, the difference appears not significant and low concentrations.

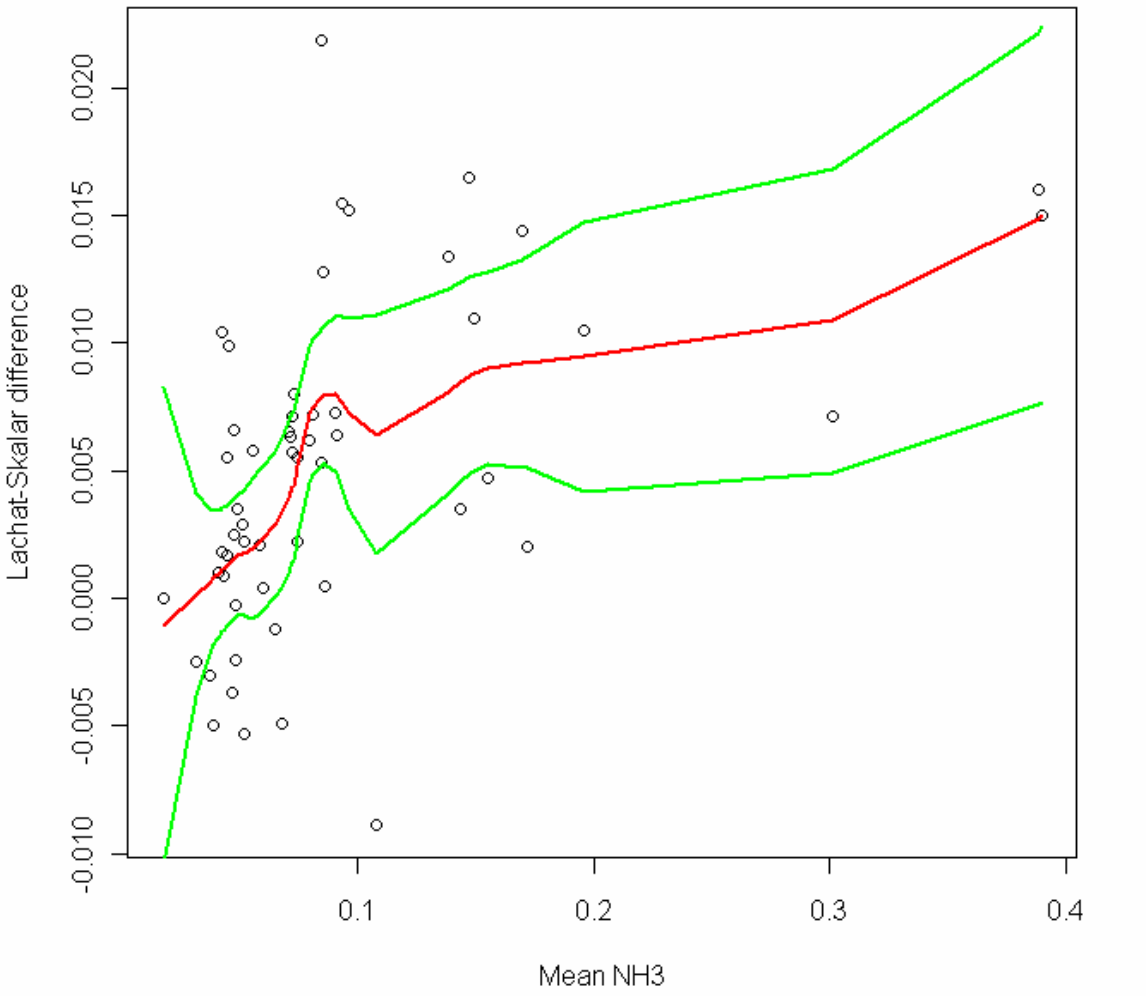


Figure 6. Plot of method differences versus mean for 2nd study.

In this study it is difficult to assess if the differences are strictly proportional to the mean because of low sample size at higher concentrations (Figure 7). If we assume strict proportionality, the adjustment factor is 0.9496054 which is less of an adjustment than for the data above.

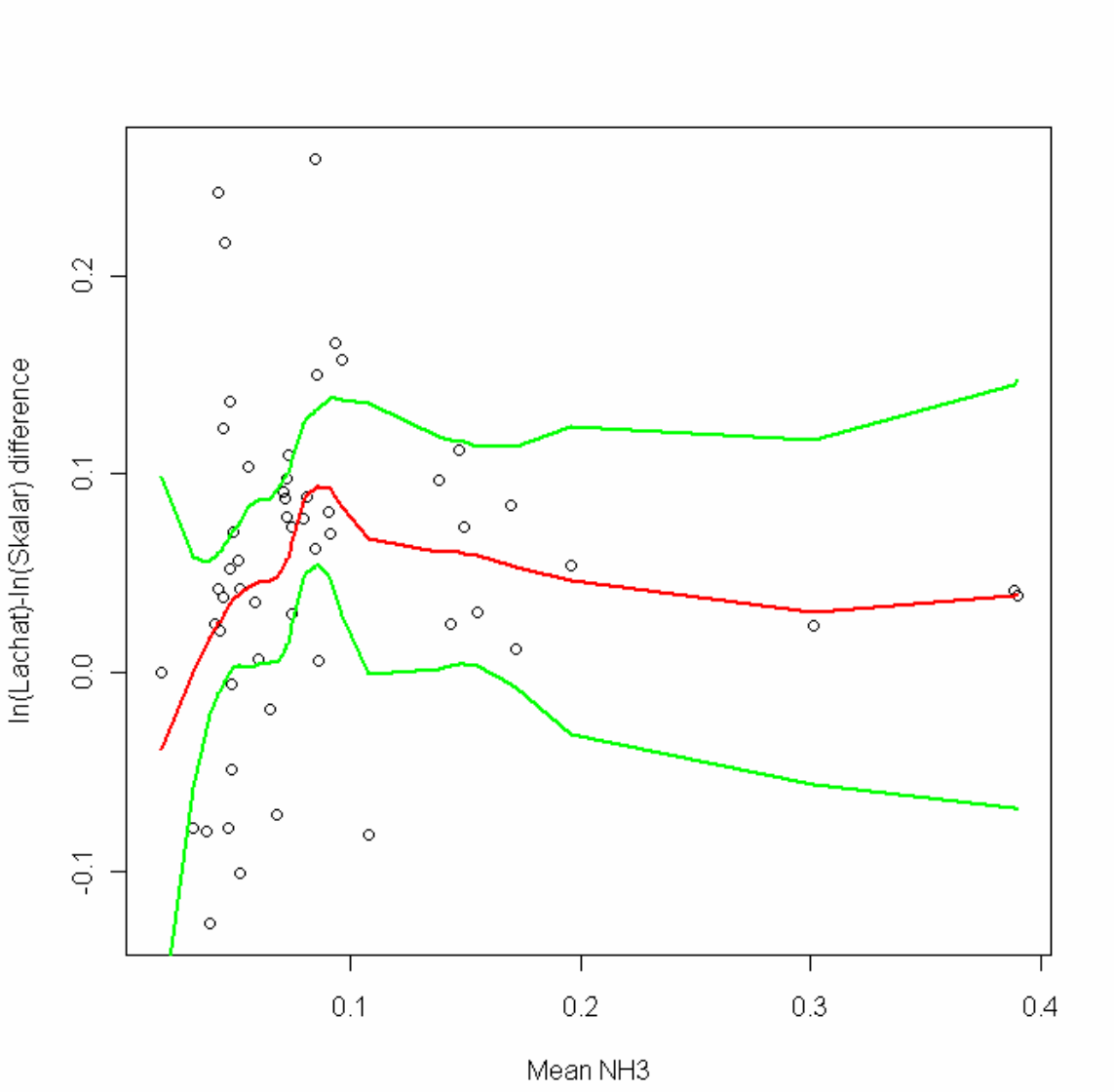


Figure 7. Differences in logarithm transformed measurements from Lachat and Skalar methods plotted against mean concentration.

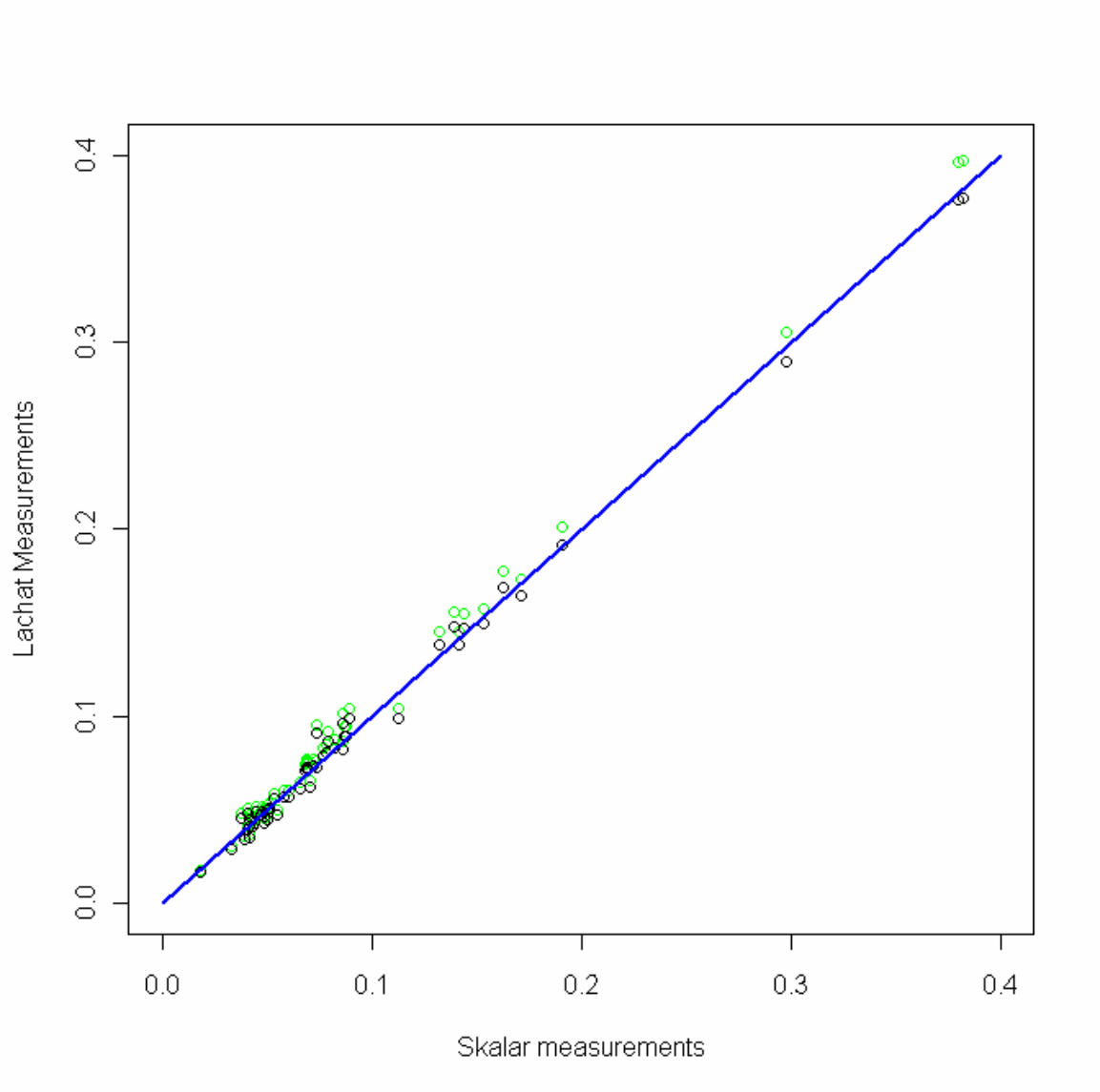
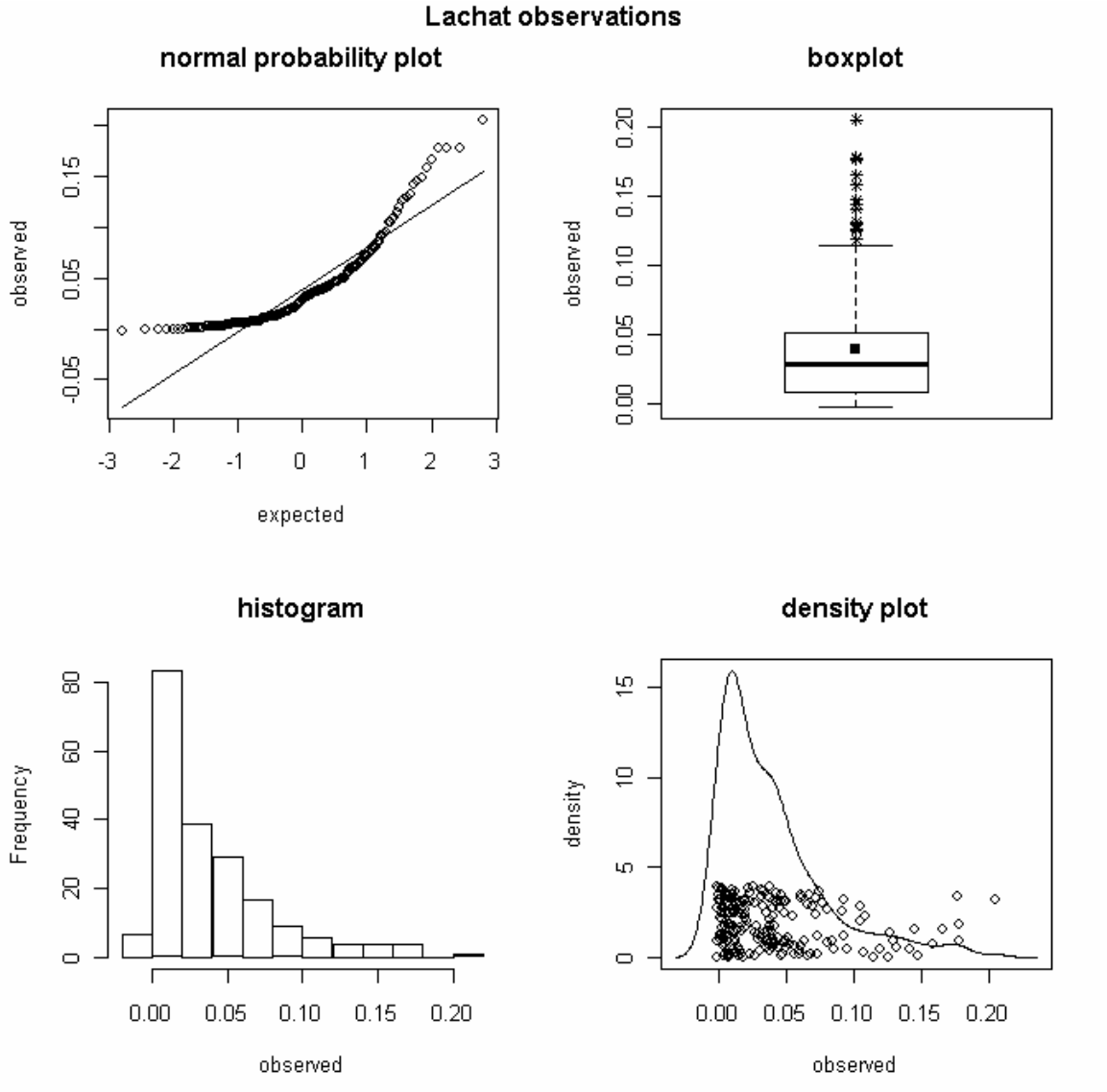


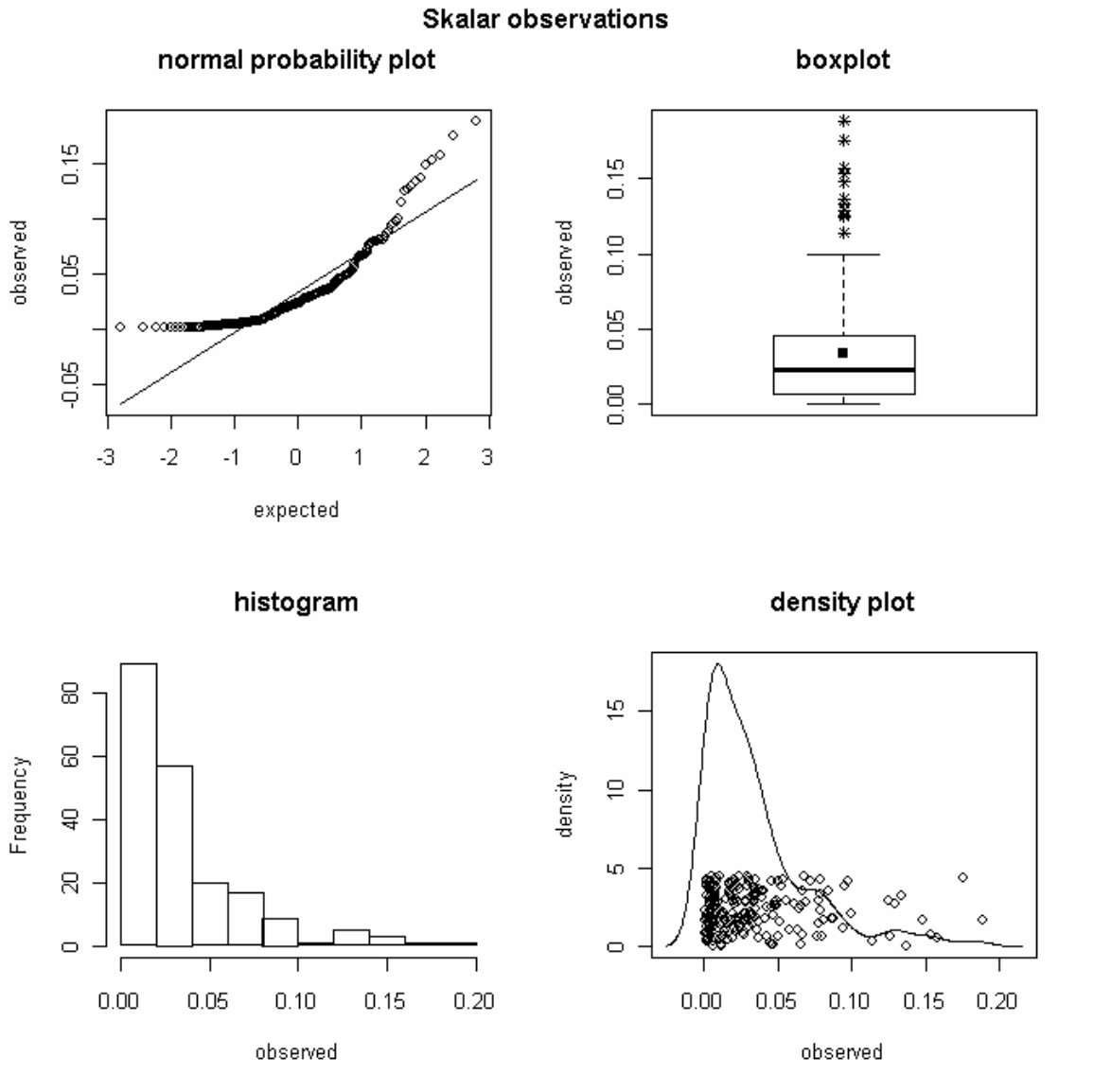
Figure 8. Skalar measurements plotted versus Lachat measurements (green circle) and adjusted Lachat measurements (black circle). The one-to-one line is shown in blue.

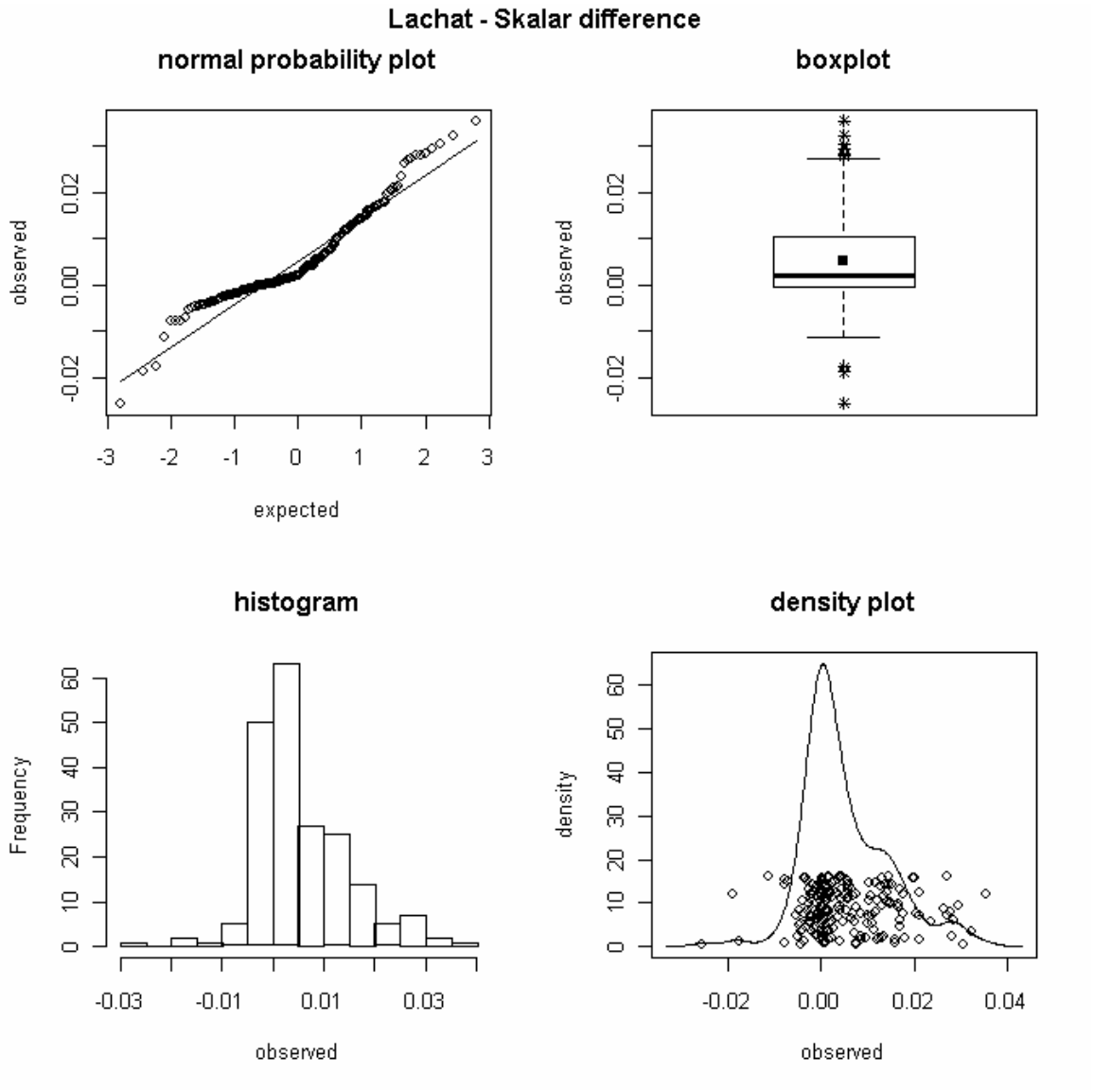
On the whole, there does not appear to be any information in this second data set that strongly refutes what was found in the first.

Distribution Plots:

The distribution plots that follow illustrate the basic distribution properties of the methods comparison data. The NH3 measurements by both methods exhibit a skewed distribution that resembles a log-normal which is typical of water quality data. The difference data have a symmetric distribution with somewhat heavier tails than expected from a normal distribution.




Skalar observations
normal probability plot
observed
expected
-3
-2
-1
0
1
2
3
-0.05
0.05
0.15
boxplot
observed
0.00
0.05
0.10
0.15
histogram
Frequency
observed
0
20
40
60
80
0.00
0.05
0.10
0.15
0.20
density plot
density
observed
0
5
10
15
0.00
0.05
0.10
0.15
0.20


Lachat - Skalar difference
normal probability plot
observed
expected
boxplot
observed
histogram
Frequency
observed
density plot
density
observed

Note to Hoffman:

From:Elgin Perry <eperry@chesapeake.net>
To:"Hoffman, Frederick (DEQ)"
<Rick.Hoffman@deq.virginia.gov>
Subject:Re: Comparison data for Ammonia Method/instrument change
Date sent: Fri, 30 Mar 2012 16:00:40 -0400
Copies to: "Johnson, Cynthia (DEQ)"
<Cindy.Johnson@deq.virginia.gov>

Rick,

I have given these results a quick review and see no major issues. I did jot down a few quick notes as I was reading. These are:

- Technical Procedure 2-600.pdf

The following are identified as interfering chemicals:
seawater calcium and magnesium
pH outside 4.0-10.0
sulfide > 2 mg/l
turbidity

No followup to assure that these did not cause problems in the comparison experiment.

The %recovery results on page 31/49 show strong autocorrelation. It is not clear why this occurs.

lack of fit test table on page 33/49 has a confusing presentation

Heteroscedastic statement not supported by stat test

split sample analysis starts page 40.
large sample size = 203
Minimum Significant Difference = 0.0012789
Because of the large sample size, this test is capable of finding that a very small difference is statistically significant. It is likely that a the significant difference found is of no practical importance.

results on page 41 and top of 42 report mean difference as postive.

T-test results show men difference as negative. Why is there a switch?

I will do a quick analysis of the data next week and let you know if issues emerge.

regards, Elgin

Subject: **FW: Comparison data for Ammonia Method/instrument change**
Date sent: **Thu, 29 Mar 2012 13:33:43 -0400**
From: **"Hoffman, Frederick (DEQ)" <Rick.Hoffman@deq.virginia.gov>**
To: **<eperry@chesapeake.net>**
Copies to: **"Johnson, Cynthia (DEQ)" <Cindy.Johnson@deq.virginia.gov>**

Hi Elgin,

Attached are some data for a method change comparison study. We would like your opinion.

Thanks,

Rick Hoffman
Chesapeake Bay Monitoring Program
VA Department of Environmental Quality
P.O.B. 1105, Richmond VA 23218
629 E. Main St. (Street Address, i.e. for Fed-ex)

804-698-4334 (Phone); 804-698-4032 (Fax, 9'th flo
Elgin S. Perry, Ph.D.
Statistics Consultant
2000 Kings Landing Rd.
Huntingtown, MD. 20639
ph. 410.535.2949