Title: Chemometrics Experiment

Alan Yu*, and Haroon Chaudhry, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08901.

Abstract:

This experiment had the first week in the lab to weigh different types of capsules, fish oil and krill oil, using two different types of scales. Using statistic t and the t-test, Krill Oil 3 between my lab partner and I were found to be statistically the same when weighed, but within the same set of samples, the t-test concluded that the weighings were different. For week two, two datasets were given where one was trace elements and pH in ice cubes from various New Jersey fast food restaurants. Using a Pearson correlation matrix, it was found that Ba was correlated with Sr with a correlation coefficient squared $R^2 = 0.99$. A plot show that the correlation coefficient squared was 0.972. The second dataset contained data of trace elements in New Jersey soils in different regions. Again, using a Pearson correlation matrix, it was found that KK was correlated with SiK with a correlation coefficient squared $R^2 = 0.93$. Another plot showed that the correlation coefficient squared was 0.858. The means, standard deviations, and the percent relative standard deviations for the trace elements in both data sets were calculated. The most concentrated trace element in ice cubes was found to be Zn with a mean of 8.87 ppb. The most concentrated trace element in New Jersey soils was found to be FeK with a mean of 990.23 ppb.

I.   Introduction

We used Chemometrics to determine the differences between weighing different dietary supplements between myself and my laboratory partner. In addition, we used Chemometrics to reduce a trace element profile of elements in contaminated NJ soils and in ice cubes.

For my pre-laboratory report, Kiefer et. al.[1] used Chemometrics to determine if passion fruit oil was adulterated from sunflower oil requiring only ~1% of sunflower oil in the mixture to be able to characterize the two oils in the FTIR spectra. I also read two additional peer-reviewed journal articles that were published in 2019 and they are summarized together with my pre-laboratory report journal article in Table 1. The first additional article that I read was by Yang et al [2] where they used chemometrics, specifically particle swarm optimization and data fusion. and spectroscopic techniques to improve the identification ability of abilities of different grades of P. notoginseng powder which is a possible adulterant. They concluded that the minimum blend ratio was 0.5% which increased the accuracy of classification to 96.97%. The second additional article that I read was by Raoufi et al [3] where they used an optimization of response surface methodology using central composite design on blood samples to more accurately determine the concentrations of Atenolol, Metoprolol, and Propranolol. They concluded that the method that was used was effective enough to determine the concentrations of these three in real applications.

Table 1. My refereed journal articles.

Corresponding Author	Date Submitted	Date Accepted	Samples	My reference number
Johannes Kiefer	August 4, 2019	September 4, 2019	Pure passion fruit oils, pure sunflower oils, and an unknown mixture of both	1
Xiaodong Yang	June 22, 2019	October 29, 2019	Blends ratios of P. notoginseng powder of different grades	2
Arastou Raoufi	July 21, 2019	October 2, 2019	Blood samples containing Atenolol, Metoprolol, and Propranolol	3

Chemometrics is where mathematical, statistical, and other methods are used to analyze the data that creates the best optimization of the information to yield the most relevant information. From

pages 27-28 of the textbook,[4] the first equation that was used was to calculate the mean of the samples where

$$\overline{X} = \frac{\sum\limits_{1}^{n} X_n}{n} \tag{1}$$

with $\overline{X}$ is the mean (average), $\sum\limits_{1}^{n} X_n$ is the sum of all the trials, and n is the number of samples done. The second equation that's used is the equation for standard deviation where

$$\sigma = \sqrt{\frac{\sum\limits_{i=1}^{n} (\overline{X} - X_i)^2}{n-1}} \tag{2}$$

with σ being the standard deviation, $\overline{X}$ is the mean (average), $\sum\limits_{i=1}^{n} (\overline{X} - X_i)^2$ is the sum of the mean minus the X_i, which is the trial of the specific term, i, squared, and n is the number of trials. After finding the standard deviation, the next equation can be used where the percent relative standard deviation where

$$\% \text{ RSD} = \frac{\sigma}{\overline{x}} * 100 \tag{3}$$

with σ being the standard deviation and $\overline{x}$ is the mean of the trials. Using another equation, confidence limits could be found where

$$\overline{X} \pm \frac{t*s}{\sqrt{n}} \tag{4}$$

with $\overline{X}$ is the mean (average), t is the student's t-distribution, n is the number of samples and s is the standard deviation. The next equation that is used is the calculation of statistic t to carry out a t-test to compare the statistics between two trials where

$$t_{experimental} = \frac{\overline{x_1} - \overline{x_2}}{s} \sqrt{\frac{n_1 n_2}{n_{1+} n_2}} \tag{5}$$

with the $\overline{x_1}$ being the mean from one test, $\overline{x_2}$ being the mean from the other test, s being the pooled variance where

$$s = \sqrt{\frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}} \tag{6}$$

with s_1 being the standard deviation of one total run, s_2 being the standard deviation of another total run, n_1 is the number of trials of one total run, and n_2 being the number of trials of another total run.

II.   Experimental

Table 2 summarizes the samples that I weighed using the Mettler Toledo Balance and their descriptive statistics.

Table 2. Dietary samples weighed and descriptive statistics.

Sample ID	Mean (g)	STD (g)	%RSD	Mean (g) from 10
FO1	0.7538	0.0140	1.86	0.7544
FO2	1.3508	0.0120	0.89	1.3508
FO3	1.8293	0.0320	1.75	1.8288
FO4	1.8167	0.0342	1.88	1.8077
KO1	0.7432	0.0124	1.66	0.7429
KO2	1.4764	0.0379	2.56	1.4766
KO3	0.6323	0.0050	0.79	0.6328
KO4	1.3598	0.0164	1.21	1.3597

Materials and Supplies:

Samples were two different types of dietary supplements, fish oil and krill oil.

Plastic Containers

Tweezers

Weighing Boats

Instrumentation:

My weighing station: Mettler Toledo (Columbus, OH), AG204, Labeled as Bal-15, Max: 210g, ±0.0001g

My lab partner's weighing station: OHaus Pioneer (Parsippany, NJ), PA214, ±0.0001g, made in China

Software:

To determine the statistics: Microsoft Excel , Microsoft (Redmond, WA), Version 1902 (Build 11328.20438).

III.   Analytical Procedure

For this experiment, only one week was required. My lab partner and I acquired four samples of oil capsules, fish and krill specifically, each and weighed each on two different balances. Each partner weighed four containers of either fish oil or krill oil, with each containing ten capsules each. The weighing procedure began as first weighing the weighing boat, taring the

boat, then weighing all ten as a group. After weighing all ten as a group, each of the capsules are weighed individually in the weighing boat. This repeated for all four containers and both partners were instructed to switch the containers they weighed and repeated this process. The second week was done without a lab partner and outside of the lab. This is where data was acquired from the instructor for trace elements and pH in ice cubes and XRF analysis on Cr waste contaminated soils.

IV.   Results and Discussions

Table 3 Summarizes my descriptive statistics versus my laboratory partner's for Krill Oil 3.

Table 3. Dietary sample's descriptive statistics.

Sample ID	Mean (g)	STD (g)	%RSD
Krill Oil 3 (Mettler)	0.6323	0.0050	0.79
Krill Oil 3 (OHaus)	0.6334	0.0054	0.85

I used the confidence limits and statistic t to carry out a t-test which is shown in the Introduction and below as equations 4, 5 and 6.

$$\overline{X} \pm \frac{t*s}{\sqrt{n}} \qquad (4)$$

with $\overline{X}$ is the mean (average), t is the student's t-distribution, n is the number of samples and s is the standard deviation. The next equation that is used is the calculation of statistic t to carry out a t-test to compare the statistics between two trials where

$$t_{experimental} = \frac{\overline{x_1} - \overline{x_2}}{s} \sqrt{\frac{n_1 n_2}{n_{1+}n_2}} \qquad (5)$$

with the $\overline{x_1}$ being the mean from one test, $\overline{x_2}$ being the mean from the other test, s being the pooled variance where

$$s = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1 + n_2 - 2}} \qquad (6)$$

with s_1 being the standard deviation of one total run, s_2 being the standard deviation of another total run, n_1 is the number of trials of one total run, and n_2 being the number of trials of another total run. Based on this determination, I conclude at the 95% confidence that my weighing method using the analytical balance from Mettler Toledo for sample Krill Oil 3 were statistically the same due to my value being ±0.0035 compared to my lab partner's being ±0.0038 using the OHaus balance. Using the statistical t and t-test, $t_{experimental}$ was found to be 0.611 which is less than 1.725 which means the weighings are the same. Using the simpler method, krill oil 1 vs krill oil 3 yields ±0.0086 vs ±0.0035. The values here are different by 0.005 but can be said that the

values are about the same, with a little bit of variance. Using the statistical t and t-test, $t_{experimental}$ was found to be 26.455 which is a lot greater than 1.725 shows that the weighings of krill oil 1 vs krill oil 3 are very different.

Part II.

Data Set 1
Table 4 summarizes descriptive statistics of trace elements and pH in ice cubes from NJ fast food restaurants.

Table 4. Descriptive statistics of trace elements and pH in ice cubes.

Element	Mean	STD	%RSD
Cu	5.40	5.90	109.40
Zn	8.87	22.53	253.93
Sr	2.85	5.44	190.98
Ba	0.81	1.35	165.98
Pb	0.04	0.05	113.08
pH	5.41	0.33	6.03

Table 5 is the Pearson's correlation matrix showing the interrelationship between the variables. I used Microsoft Excel (Microsoft, Redmond, WA) to compute the variables in this table.

Table 5. Pearson's correlation matrix ice cube dataset.

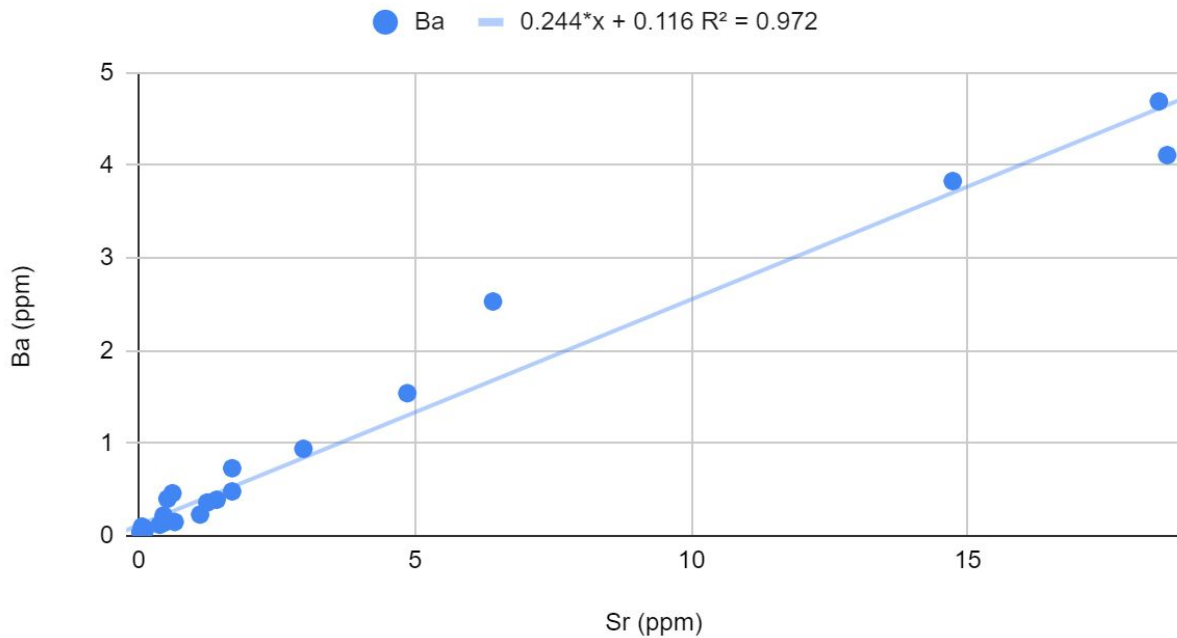|	Cu	Zn	Sr	Ba	Pb	pH
Cu	1					
Zn	0.15	1				
Sr	0.62	0.09	1			
Ba	0.60	0.10	0.99	1		
Pb	0.17	0.85	0.09	0.11	1	
pH	0.14	0.07	0.22	0.27	0.18	1

Figure 1. Plot created to display the correlation equation of Ba vs Sr.

From Table 5, I conclude that element Ba is correlated with element Sr. Therefore I can simplify the dataset by substituting element Ba with the following equation 7:

$$Y = 0.244x + 0.116 \qquad\qquad (7)$$

where Y is the concentration of Ba in ppb, 0.244 is the slope, the value x is the concentration of Sr in ppb and from the correlation matrix, the correlation coefficient squared is $R^2 = 0.99$ and explains 99% of the variance between Sr and Ba. From the plot however, the correlation coefficient squared is 0.972, which is close to the stated value in the matrix, but using the equation would explain 97.2% of the variance. A close data sample is between Pb and Zn, but since the correlation squared value is 0.85, it is sufficient to exclude it from consideration.

Data Set 2

Table 6 summarizes descriptive statistics of trace elements in NJ soils.

Table 6. Descriptive statistics of trace elements in NJ soils.

Element	Mean (ppb)	STD (ppb)	%RSD
MgK	7.51	5.38	71.65
AlK	31.40	24.22	77.11
SiK	135.58	141.80	104.59
SK	19.21	36.29	188.94
RhL	29.40	10.50	35.72
KK	17.75	20.95	118.05
CaK	654.28	493.35	75.40
TiK	20.43	14.39	70.46
VK	8.59	8.18	95.32
CrK	225.92	271.02	119.96
MnK	11.99	7.49	62.48
FeK	990.23	607.96	61.40
NiK	5.84	4.62	79.14
CuK	3.01	2.53	84.04
ZnK	6.98	5.90	84.52

Table 6 is the Pearson's correlation matrix showing the interrelationship between the variables. I used Microsoft Excel (Microsoft, Redmond, WA) to compute the variables in this table.

Table 7. Pearson's Correlation Matrix of trace elements in NJ soils.

|	MgK	AlK	SiK	SK	RhL	KK	CaK	TiK	VK	CrK	MnK	FeK	NiK	CuK	ZnK
MgK	1														
AlK	-0.38	1													
SiK	-0.61	0.25	1												
SK	-0.13	-0.14	-0.15	1											
RhL	0.43	-0.35	-0.63	-0.16	1										
KK	-0.63	0.30	0.93	-0.07	-0.61	1									
CaK	0.31	-0.41	-0.73	0.24	0.65	-0.68	1								
TiK	-0.20	0.30	0.56	-0.26	-0.41	0.59	-0.76	1							
VK	0.58	-0.09	-0.47	-0.22	0.40	-0.45	0.06	0.11	1						
CrK	0.42	-0.11	-0.49	-0.14	0.24	-0.49	0.01	-0.01	0.47	1					
MnK	0.50	-0.17	-0.25	-0.19	0.23	-0.25	-0.04	0.20	0.82	0.29	1				
FeK	0.70	-0.12	-0.43	-0.26	0.35	-0.43	-0.06	0.24	0.82	0.55	0.77	1			
NiK	0.56	-0.29	-0.59	-0.01	0.50	-0.57	0.45	-0.25	0.72	0.25	0.78	0.54	1		
CuK	-0.15	0.09	0.25	-0.01	-0.21	0.26	-0.18	0.16	0.22	-0.21	0.42	-0.04	0.46	1	
ZnK	0.10	0.19	-0.11	0.37	-0.10	-0.04	-0.13	0.07	0.31	0.07	0.31	0.26	0.28	0.42	1

KK (ppm) vs. SiK (ppm) in various New Jersey Soils
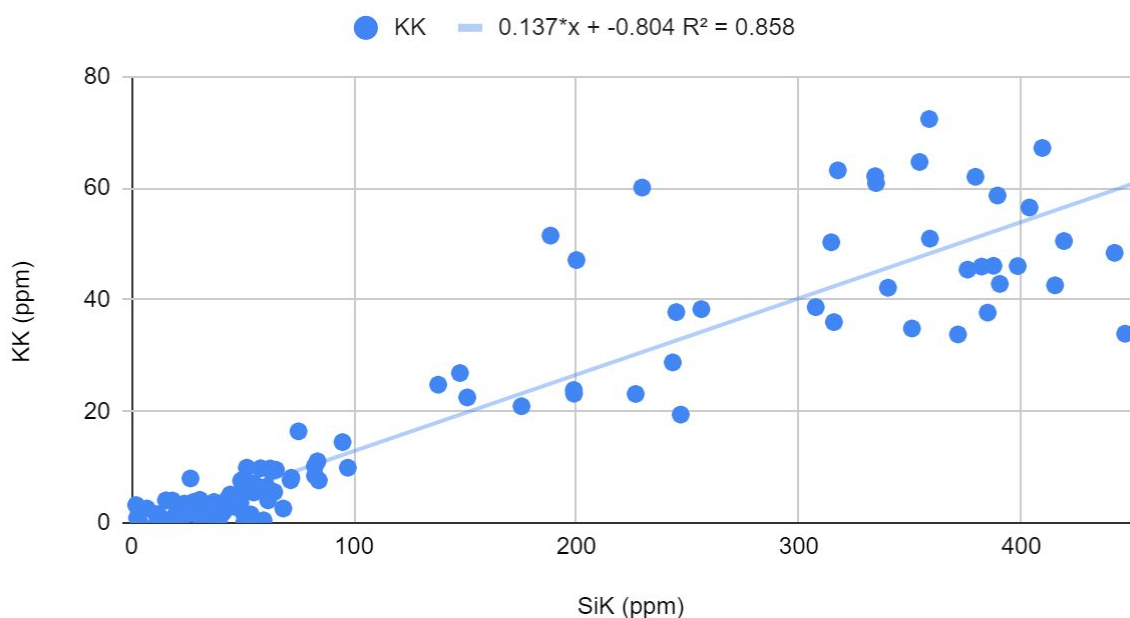
Figure 2. Plot created to determine the correlation equation between KK and SiK.

From Table 7, I conclude that element KK is correlated with element SiK. Therefore I can simplify the dataset by substituting element KK with the following equation 8:

$$Y = 0.137x - 0.804 \qquad (8)$$

where Y is the concentration of KK in ppb, x is the concentration of SiK and 0.137 is the slope. and from the correlation matrix, the correlation coefficient squared is $R^2 = 0.93$ and explains 93% of the variance between SiK and KK. The correlation coefficient squared from the plot is 0.858 however, so using this equation can explain the85.8% variance between the two elements. No other data correlations had a correlation coefficient squared value that was greater than 0.90 and are disregarded.

V.   Conclusions

Between my lab partner and I, the sample Krill Oil 3 weighings were found to be the same from a t-test, but between the same set of capsules, Krill Oil specifically, the measurements were concluded to be statistically different.

Based on the Pearson's correlation matrix, it can be concluded that Ba and Sr are correlated together in the composition of ice cubes from fast food restaurants, with an equational relationship of $Y = 0.244x + 0.116$ where Y is the concentration of Ba in ppb and x is the concentration of Sr in ppb with a correlation coefficient squared of $R^2 = 0.99$ from the correlation matrix. Figure 1 shows the correlation coefficient squared of 0.972 which can be used to explain 97.2% of variance between Ba and Sr.

Using the Pearson's correlation for the second data set, it was found to be that KK is correlated with SiK in the composition of soils from different places in New Jersey. The equation that correlates the two elements is $Y = 0.137x - 0.804$ where Y is the concentration of KK in ppb and x is the concentration of SiK in ppb with a correlation coefficient squared of $R^2 = 0.93$ from the correlation matrix. However, from the plot itself, the correlation coefficient squared of 0.858 which can be used to explain 85.8% of variance between KK and SiK.

From data set 1, it was found that Zn has a mean concentration of 8.87 ppb. From the data in data set 2 collected, the mean of FeK was 990.23 ppb. This was the highest mean in the dataset that was provided with CaK with 654.28 ppb coming in second as the highest concentrated element in New Jersey soils.

For New Jersey soils, any trace elements that is high in concentration will be toxic to humans. A standard for trace elements is about 1000ppb or 1ppm. From the data set using the mean and standard deviations, elements CaK and FeK are extremely high in concentration. The upper end of CaK can be around 1,147.63ppb and for FeK, it can be around 1,598.19ppb. These two are considered toxic because of their amounts being larger than what a trace element definition can be.

VI. References

1. Johannes Kiefer, Anja I. Lampe, Stefano F. Nicoli, Massimo Lucarini, and Alessandra Durazzo, Identification of Passion Fruit Oil Adulteration by Chemometric Analysis of FTIR Spectra. Molecules 2019, 24(18), 3219.

2. Xiaodong Yang, Jie Song, Lin Peng, Lutao Gao, Xuwen Liu, Lin Xie, and Guanglin Li, Improving identification ability of adulterants in powdered Panax notoginseng using particle swarm optimization and data fusion. Infrared Physics & Technology 2019, Volume 103 (103101).

3. Arastou Raoufi, Mahmoud Ebrahimi, and Mohammad Reza Bozorgmehr, Application of response surface modeling and chemometrics methods for the determination of Atenolol, Metoprolol and Propranolol in blood sample using dispersive liquid–liquid microextraction combined with HPLC-DAD. Journal of Chromatography B (2019), Volume 1132 (121823).

4. Undergraduate Instrumental Analysis, James W. Robinson, Eileen M. Skelly Frame, George M. Frame II, CRC Press, 2014, 7th edn.