THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF SUPPLY CHAIN AND INFORMATION SYSTEMS

ESTIMATING YIELD DISTRIBUTIONS FROM TRUNCATED DATA USING LINEAR REGRESSION MODELS

ERIK ZALEWSKI

SPRING 2016

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Supply Chain and Information Systems
with honors in Supply Chain and Information Systems

Reviewed and approved* by the following:

Saurabh Bansal
Assistant Professor of Supply Chain Management
Thesis Supervisor

John Spychalski
Professor Emeritus of Supply Chain Management
Honors Adviser

*Signatures are on file in the Schreyer Honors College

ABSTRACT

This thesis focuses on a problem in the context of the commercial seed industry. Specifically, it investigates how seed producers can estimate the distribution of production yields when only truncated data from some trials is available, i.e., the numerical values are observable for some but not all data points. There are a few different ways to combat this issue. One is to take the data points available, assume that they constitute the entire data set, and then use them to estimate the mean and standard deviation. Another approach is to acknowledge the presence of other unobserved data points and estimate the mean and standard deviation by assigning a weight to each observation available. The thesis describes a mathematical development for the second approach, and then shows that this approach is superior over the first approach. Using illustrative example, the increase in dollar amount for a representative problem in the commercial seed industry is also calculated.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Dr. Saurabh Bansal and Dr. John Spychalski for their help in the development and completion of this thesis.

Chapter 1

Introduction

Supply chain contracts are common in business and involve two or more parties negotiating on the prices and quantities of good being delivered. The focus in this thesis is on contracting situations in which the production process is subject to a random yield. Random yield means that the output from a production system is unknown. This uncertain output creates a supply risk in the supply chain. As a result, when designing the contract, there needs to be a determination from the two sides about how and to what proportion the yield risk will be divided. A scenario, and one that this thesis could be helpful for, is if a supplier is responsible for providing a fixed amount to the buyer and bears the complete yield risk. From the supplier’s perspective she should charge a premium for bearing the risk. From the buyer’s perspective the premium paid should be commensurate to the risk. Therefore it is necessary for the buyer to understand how uncertain the yield is. With this information, buyers can better negotiate contracts with suppliers. In this thesis we will focus on quantifying the yield uncertainty in terms of its probability distribution.

When trying to determine the probability distribution of the yield, it is usually necessary to rely on past data. But sometimes all past data may not have numerical values. Our focus will be on the case when a number of observations were made for the yield, but only a portion of these observations actually have a particular numerical value attached to them. For example, we consider the problem faced by an agribusiness firm that contracts with contract-farmers to produce seeds. A common contract in this domain is that if the total production from a field exceeds the amount requested by the firm the firm does not observe the yield obtained by the contract-farmer. But if the total production is lower than the ordered quantity then the firm will be able to deduce the yield obtained in the field from the amount of seed delivered by the contract-farmer. Say a firm entered into such contracts with 30 contract-farmers for specific amounts. In 19 of these cases the yield were low, the contract-farmers were not able to meet the firm's demand and transferred their entire output to the firm. The firm deduced the yield values for these 19 observations. For the remaining 11 cases the firm knows that the yields exceeded a threshold but does not know their true values. This scenario is known as having truncated data. For a side of the contract negotiations such as the buyer, not having that data available can make determining yield estimates a truly tough task.

There are a few ways to construct probability distributions from truncated data. One approach is to take the observations that are available, say the 19 data in the context discussed above, and then use these data to determine the mean and standard deviation. Another method, and the one in which this thesis will attempt to prove is better, is to reference the fact that there is a larger data set, the 30 observations, by applying specific weights to the known numerical observations such that the weights take into account the total sample size of 30 and the number of available values 19. In this way, the missing data are not ignored, and instead, are taken into account. Over the course of the thesis, an analytical method will be discussed for the latter approach and a numerical benefit will be discussed to show how helpful recognizing the non-numerical observations can be. Finally, a numerical study will quantify the likely monetary impact of using approach in industry setting.

Business Application

In the previous section of the introduction, the general idea of the thesis was explained. We now expand more on the industry context that motivated this research. This research is motivated by the commercial seed industry. In this industry, a company, say Company A, sell seeds to farmers so that they can grow crops on their land. However, the seeds themselves were not made by Company A, instead they were grown by suppliers who sell their seeds to Company A. Contracts between a firm like Company A and the supplier are designed such that Company A pays a contracted price to the seed supplier in exchange for the desired amount. Contracts are done annually. The negotiated price typically depends on the yield that suppliers can obtain from their land. However, not all parties of the agreement typically have the yield data from prior years.

As an illustration suppose Company A ordered 1000 pounds of seed from their particular supplier of seeds. When it came to picking up the order, the supplier reveals that they were only able to produce 700 pounds of seed from say 10 acres of land they have. As a result, Company A can clearly mark that the yield for that year was 70 pounds per acre of seed for that particular supplier. However, in the event that the production was above the ordered amount of 1000 pounds, then Company A would not have an idea of what the actual yield was. Now, take that example for the one-year and expand it to cover a ten-year span. There is the possibility that Company A may only know the yield for six out of the ten years. The company is then at a disadvantage when it comes to negotiations because they do not have all of the historical data while the supplying farmers do. At that point, in order to predict the yield distribution, Company A could use the two approaches to estimate the yield distribution described earlier in the thesis. They could take the yields from the six years with the known data and attempt to do something as simple as take the average of the outputs and use that as the predicted yield for the coming year. Or Company A could take the second approach of taking into account that there is ten years worth of data, and although only six years carry a numerical value, there is still worth in acknowledging that there were four years that exceeded the prescribed order amount. The challenge is to translate this acknowledgment into a mathematical approach. The goal of this thesis is to show how this can be done.

Methods to be used

I will use a number of different analytical tools in the thesis. These tools include software include Microsoft Excel, in which both regression analysis and the Solver function were used, and Matlab, which is used thoroughly to show the difference between the two main approaches to create a yield estimate. Sections later in the thesis will describe how each of these programs can aid in accomplishing the goal of testing the hypothesis. Results will be listed in each section and analyzed to determine their meaning and importance to the overall task at hand.

Chapter 2

Literature Review

The main academic work that we draw upon is a journal article by E.H. Lloyd from Imperial College, which is entitled “Least-Squares Estimation of Location and Scale Parameters Using Order Statistics”. Within the paper is an important formula that will be used in testing the efficacy of the two approaches, ignoring non-numerical, unknown data or embracing it. The formula as listed in the paper is as follows,

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{p}' \; \boldsymbol{\Omega}\boldsymbol{p})^{-1}\boldsymbol{p}' \; \boldsymbol{\Omega}\boldsymbol{Y} \qquad (1)$$

with the vector $\boldsymbol{\theta}$ of mean and standard deviation (Lloyd, 1952), and $\boldsymbol{Y}$ is the vector of the observations for which data are available such that all data are below a certain threshold value. The portion of the formula before the $\boldsymbol{Y}$ variable denotes the weights assigned to each observation, that has a numerical value, in order to estimate the mean and standard deviation. If one goes back to the example within the introduction, $\boldsymbol{Y}$ would list all of the yield values that are below the order quantity of say 100 pounds per acre. The symbol $\boldsymbol{\Omega}$ denotes a variance-covariance matrix of a master data set that will discussed in more detail shortly. We will also define $\boldsymbol{p}'$ shortly with more details. Finally, we will compare the estimates of the mean and standard deviation obtained from this approach with the estimates obtained from the conventional formulas used on the data for which numerical values are available.

Other scholarly articles have shown that using additional information can be beneficial to forecasting. In the journal article “Analysis of Perishable-Inventory Systems with Censored Demand Data”, it is noted the tradeoffs that are involved when trying to use inventory to hinder censored data. The authors write *“...although stocking more units* *increases the likelihood of overstocking and therefore increases the cost of the current period, it reduces the chance of demand being censored, and hence improves the accuracy of demand estimation for future periods*" (Lu, Song, Zhu, 2008). Here, the additional information is obtained by buying more inventories. Inventories that do not get stocked out provide valuable information for the future because then there is less censored data. In the article, it notes how there is an "information benefit", which is the reduction in costs and "better future inventory decisions" that come as a result of having elevated inventories (Lu, Song, Zhu, 2008). From the work done in their paper, the results showed that "*the optimal solution uses the information to improve the demand-distribution updating*" and "*under the optimal policy, the higher the inventory level is, the more the reduction of the future cost (or equivalently, the larger the informational value) will be*" (Lu, Song, Zhu, 2008). Thus, in this case, it is noted how a company could obtain additional information by increasing inventory levels.

Another article that shows the benefit of information is entitled "Repeat tourism in Uruguay: modelling truncated distributions of count data". The issue that the article discusses is that in surveys of visitors to Uruguay, a count method was used to record how many visits a tourist has made to the country in all. However, there were specific categories for visits 1 through 5, but after five visits, all tourists were grouped together into a single, large category. Due to this counting procedure, there was accurate data for the number of tourists who visited the country five times or fewer but not more, e.g., 70 times. While this clustering was done in part because those who have traveled to a country numerous times have difficulty remembering the specific number of visits (Brida, Pereyra, Scuderi, 2012), nevertheless, this resulted in a truncated distribution. To estimate the underlying distribution of the number of visits, the authors used a quantile regression. In the regression, socioeconomic characteristics and costs incurred to go to the location were used in part of the model to estimate the distribution. Overall, this article shows again the benefit of using additional information, in this case socioeconomic characteristics, in finding out more about truncated data.

Finally, an article from Timothy J. Lowe and Paul V. Preckel entitled "Decision Technologies for Agribusiness Problems: A Brief Review of Selected Literature and a Call for Research" summarily goes through a litany of issues that the agribusiness industry faces. One anecdote within the article mentioned how another group of researchers used linear programming for optimization, but not just for determining operational actions that would maximize profit. In what was called "goal programming", a team worked with Senegalese farmers to take into account what goals they had from farming, whether it was growing enough food for their families or making more money so that it could be invested in livestock (Lowe, Preckel, 2004). After conducting a survey with farmers where they discussed their goals in farming, weights were made for the "goal program", thus producing an effective model (Lowe, Preckel, 2004). Thus, again it is shown that including additional information and attaching it to a model has a beneficial effect on the results.

To the best of our knowledge, using truncated data for yields to estimate agribusiness yields and then applying these distributions to production planning problems has not been discussed before.

Chapter 3

Analytical Models

Let μ, σ denote the mean and standard deviation, respectively, of the probability distribution. We take a sample of *i=1,2,…,n* observations, which we denote as X_i. Next, we arrange these *n* observation in the ascending order and denote the values as $Y_1, Y_2, \dots, Y_n$. We focus on the case when only *k* of these values are available. In the context of the seed industry this situation occurs when the firm has placed orders for say 100 bags per acre from *n=10* farmers. A total of *k=6* farmers obtained yields that were less than 100, and so the firm has these data available, but does not have the data available for the remaining 4 instances.

Benchmark

The benchmark approach for estimating yields is to take the observations below a the threshold value, and then use those values to estimate the mean and standard deviation for future outputs, as

$$\hat{\mu} = (\sum_{t=1}^{k} Y_t)/k,$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum_{t=1}^{k} Y_t^2 - \hat{\mu}^2}{k-1}}$$

According to this thesis's hypothesis, this approach is less accurate than the second approach; an order statistics based computation for mean and standard deviation.

Order Statistics Based Computation for Mean and Standard Deviation

The order statistics based computation is an alternative approach to estimate mean and standard deviation from truncated data. We first discuss how order statistics can be used to estimate mean and standard deviation when all numerical values are available. Then we discuss how this approach is useful when the numerical values are observable only for a few data in the observation set. We first define the standardized values $V_r; r = 1, \dots, n$ as

$$V_r = (Y_r - \mu)/\sigma \qquad (3)$$

$$V_1 \leqq V_2 \leqq \cdots \leqq V_n$$

Subsequently we define α_r as the average value of the *rth* ordered observation in a sample of size *n:*

$$E(V_r) = \alpha_r \qquad (4)$$

We also define the covariance between the observations *i,j* (i.e., between V_i, V_j) as

$$Var(V_i, V_j) = \omega_{ij} \qquad (4)$$

and denote the variance covariance matrix for all V_i as **ω.**

Lloyd (1952) showed that from the ordered observations $\boldsymbol{Y}$, the variance covariance matrix **ω,** and the matrix $\boldsymbol{p} = [\mathbf{1}\ \boldsymbol{\alpha}]$**,** the mean and standard deviations are estimated as:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{p}'\boldsymbol{\Omega}\boldsymbol{p})^{-\mathbf{1}}\boldsymbol{p}'\boldsymbol{\Omega}\boldsymbol{Y} \qquad \mathbf{(5)}$$

where $\boldsymbol{\Omega} = \boldsymbol{\omega}^{\mathbf{-1}}$.

Next we discuss how this theory can be used to estimate the mean and standard deviation when only *k* of *n* data have numerical values. We will specifically focus on the case when these *k* observations are below a threshold T. This may be the case when the firm places an order for say 1000 bags of seed, and is aware that if the contract-farmers' yield is below T, they would not be able to provide complete demand. Of *n* farmers that the firm signs the contract with, *k* of them eventually observe the yield lower than T and hence the firm is aware of the yields in these instances.

When only *k* of *n* observations are available, the firm will continue to use formula (5) above, with the difference that the matrix sizes will be equal to *k*. Specifically, ***p*** will be of size *2xk* composed of the first *k* columns of the original matrix, and **Ω** will be *kxk* and will be composed of the first *k* rows and columns of the original matrix. Intuitively by using these elements we are acknowledging that our available *k* values are from a set of *n* values.

Chapter 4

Simulation Study

We next discuss a simulation study conducted to compare the performances of the two approaches for estimating the mean and standard deviation for a normal yield distribution. The focus on the normal distribution is motivated by its widespread use in the agribusiness domain.

We used Matlab for the study. The set up for the Matlab code involved first setting up a master data set HW1 with random numbers of mean 0 and standard deviation of 1. This master data set comprised of 100,000 sets of 30 numbers, each generated independently. Each set of 30 numbers was put in ascending order. Next, the following code is used

```
Cv = cov(HW1);
alphafor10=mean(HW1,1);
```

in order to create a variance-covariance matrix for the master data set, and to calculate the alphas for each column of data. In this scenario, since we are generating sets of 30 numbers repetitively there are 30 alphas, and the matrix **Ω=**Cv is of size *30x30.* This represents our training data set.

To create the testing data set, the set-up first starts with setting aside space for 1000 iterations, and then creating a set of 30 randomly generated numbers with mean 100 and standard deviation 5 (alternations will also be done with this number). These numbers are then sorted as in the training set. A countif function is then set-up with the code in order to calculate what numbers are below the designated threshold value. The code below is for the threshold value T=100, and we will report the results for different threshold values.

```
HW2 = sort(HW2,2);
c=0;
for d = 1:size;
        if(HW2(1,d)<100)
            c=c+1;
        end
end
```

The variable “c” is the count of the numbers below the threshold value (in this case 100). With those steps, the data is now set up to complete the two different approaches.

Computation of Mean and Standard Deviation Using Benchmark 1

To compute the mean and standard deviation, a simple mean and standard deviation was needed for the truncated data. The code for this function was the following:

```
y=HW2(1:c);
temp = zeros(1, 2);
temp(:,1) = mean(y);
temp(:,2) = std(y);
thetasnaive(ex,:) = [temp(:,1)  temp(:,2)];
```

Here, “y” is the set of all observation that are below the threshold value. We take the simple mean and standard deviation of these numbers and denote them as “thetanaive”, a matrix with estimations of the mean and standard deviation for this scenario.

Computation of Mean and Standard Deviation Using Order Statistics Based Approach

In order to compute the mean and standard deviation for the order statistics based approach, we first determined how many observations in each of the set of 30 observations were below the threshold. Suppose that 15 numbers were below the threshold value, then c would be 15, and a 15 x 15 variance-covariance matrix would be taken from the master 30 x 30 matrix. Finally, only the alphas were taken for those columns below the threshold. The following code was used to perform this computation:

```
    clear onesv;
    Covar = Cv(1:c,1:c);
    alpha=alphafor10(1:c);
    ctemp=c;
    for i=1:ctemp
        onesv(1:i)=1;
    end
```

Once the appropriate $\boldsymbol{\Omega}$ and $\boldsymbol{p}$ were available we used formula (5) to estimate the mean and standard deviation, using the following code:

```
Sigma=inv(Covar);
p=[onesv'  alpha'];
pdash=transp(p);
theta=(inv(pdash*Sigma*p))*pdash*Sigma*y';
thetasorder(ex,:) = transpose(theta);
end
```

In the next section, we compare the mean and standard deviations obtained from the two approaches.

Chapter 5

Results for Estimation of Mean and Standard Deviation

The simulation study provided 1000 estimates of mean and standard deviation from the benchmark approach and using our approach. We next compared these values against the true values to determine which approach was more accurate. To show the results in a clear manner, graphs will be used to demonstrate the difference between the estimates obtained from the two approaches. Five thresholds were chosen to test the approaches: 90, 100, 105, 125, and 150 with the intent of showing if threshold values affect the performance of the approaches at all. Two graphs were made for each of these thresholds; one showing how the estimates' means performed against the target mean, and one showing how the estimates' standard deviations performed against the target standard deviation.

Figures 1(a) and (b) show the results for the threshold value of 90. Figure 1(a) shows that as the standard deviation of the true distribution increases, the order statistics based approach continues to provide high quality estimates of the standard deviation. But the performance of the naïve approach deteriorates continuously, creating a large gulf between the performances of the two approaches.

Figure 1(b) shows that this trend persists for the estimate of the standard deviation.

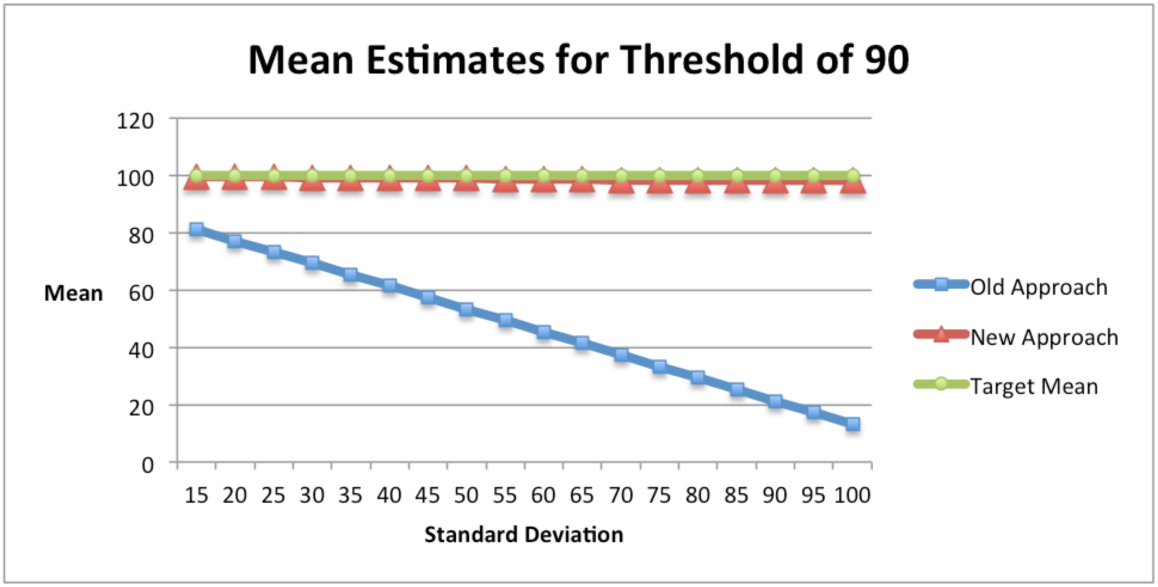

Figure 1. Mean Estimates for Threshold of 90

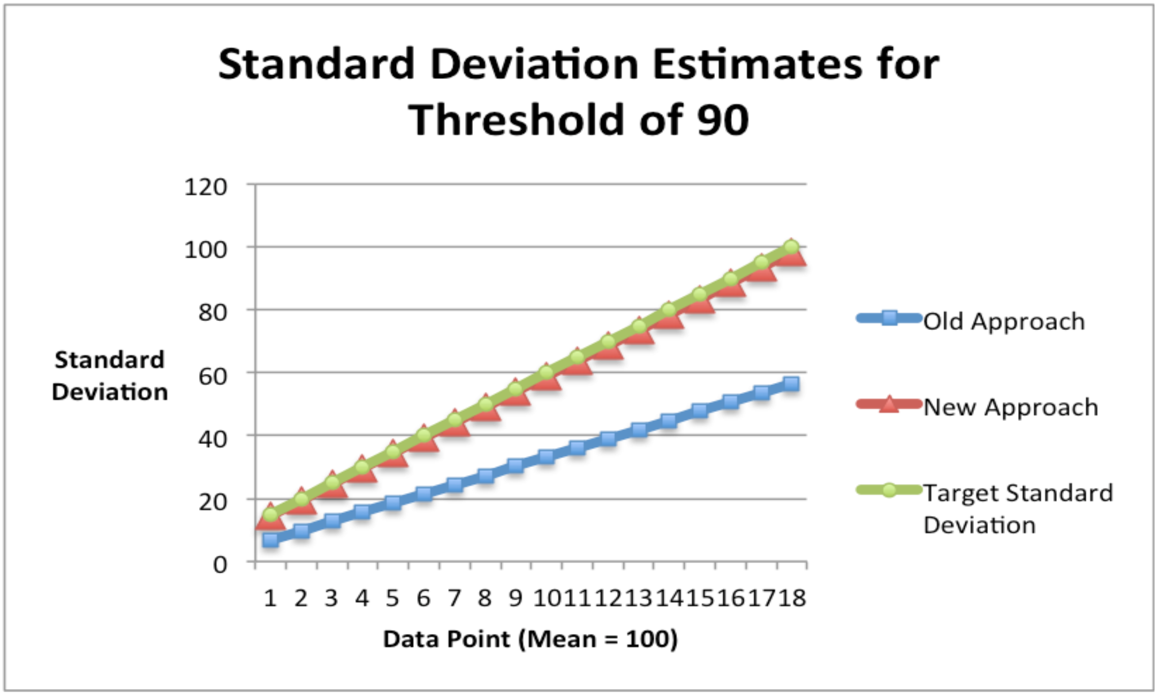

Figure 2. Standard Deviation Estimates for Threshold of 90

Next, the threshold value of 100 was tested:

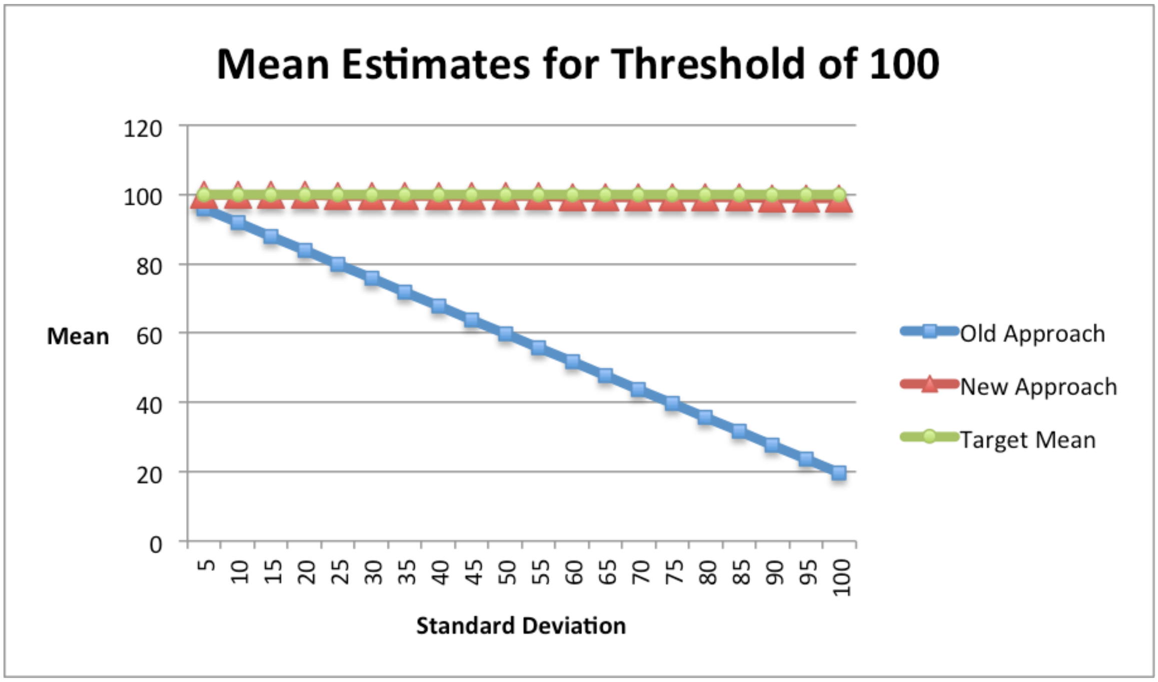

Figure 3. Mean Estimates for Threshold of 100

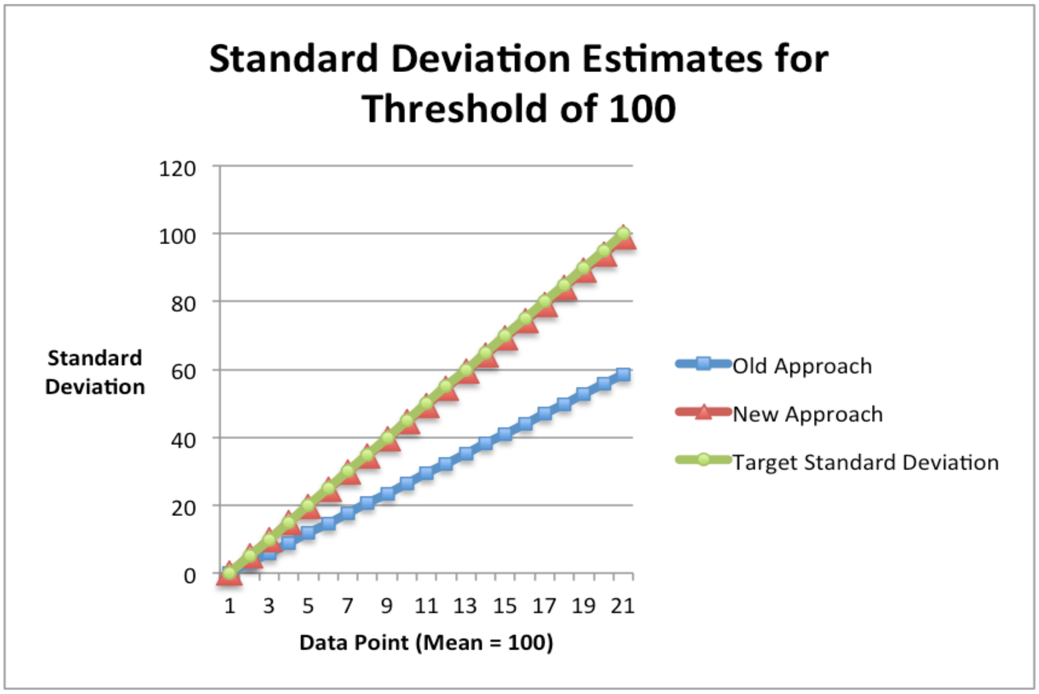

Figure 4. Standard Deviation Estimates for Threshold of 100

The threshold of 100 graphs shows much of the same determination. The new approach performs extremely well while the old approach of using truncated data does not reach the target by a significant margin.

Figures 5 and 6 show the same trend to persist for the threshold value of 105.

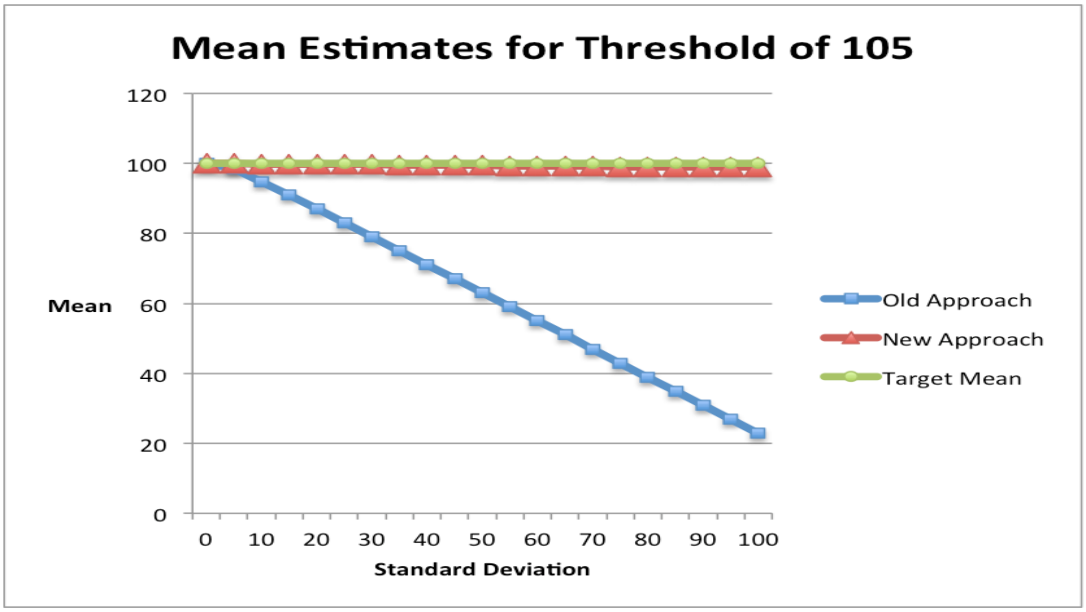

Figure 5. Mean Estimates for Threshold of 105

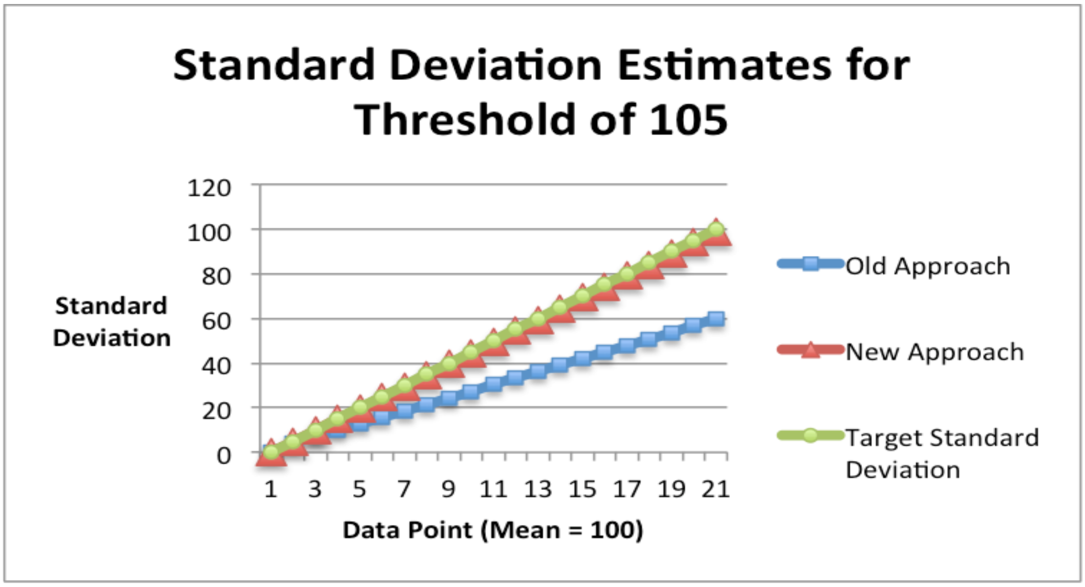

Figure 6. Standard Deviation Estimates for Threshold of 105

As before, the mean estimates for the non-truncated data, ordered estimate approach reveal results that are extremely close to the targets, while the old approach is lacking in performance. Do note the improved performance for the truncated data approach as the threshold increases. Although still not performing better than the ordered estimate approach, the old approach sees a small improvement between graphs.

Figures 7 and 8 show the results for the threshold of 125.

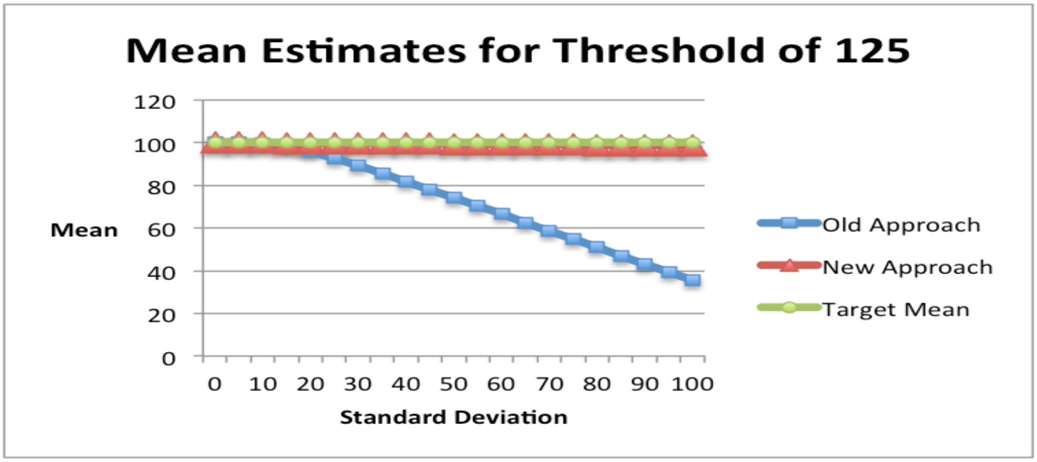

Figure 7. Mean Estimates for Threshold of 125

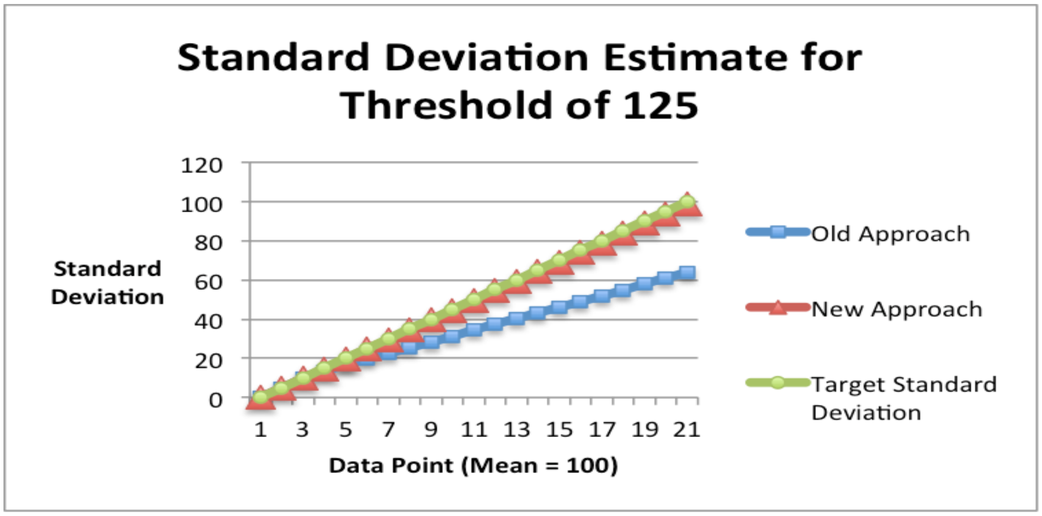

Figure 8. Standard Deviation Estimates for Threshold of 125

Although the overall difference is not that significant between this pair of graphs and the previous sets, there are some noticeable changes within the graphs. Notice in the mean estimate graph, at the beginning of the x-axis, from standard deviations 0 to about 15, the two approaches are actually very similar and perform well. Only when the standard deviation continues to increase does the new approach outshine the old one. Meanwhile, in the standard deviation estimate graph, the old approach continues to do better, even having a respectable trajectory. As seen with the mean graph, at the first few standard deviations, the truncated data method keeps pace with the target. With higher standard deviations, the method eventually falls apart compared to the ordered method, however.

Finally, the last threshold to be tested was of 150.

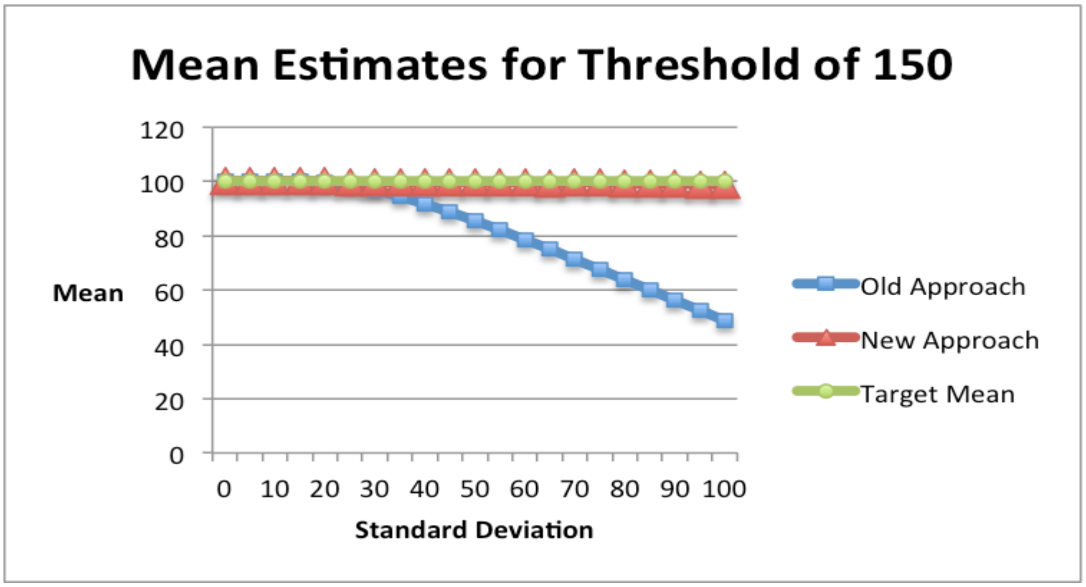

Figure 9. Mean Estimates for Threshold of 150

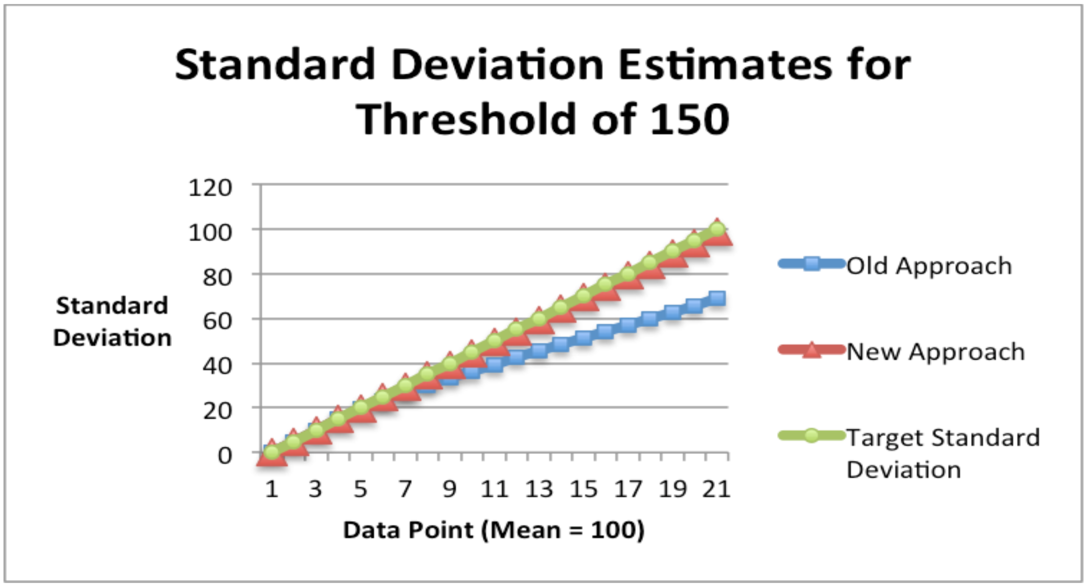

Figure 10. Standard Deviation Estimates for Threshold of 150

Here, the mean and standard deviations of the truncated approach continue to perform better, matching the targets for a longer time.

To sum the findings of this section the ordered estimate approach, the “new approach” in the graphs, was overwhelmingly the better method to use to predict the mean and standard deviation of the yield distribution. Each instance, no matter the threshold value, returned the same outcome – using an ordered estimate approach to predicting yields is the preferred method.

With regards to the threshold value, a trend that one can see from the graphs is that the benchmark truncated data approach performed better as the threshold value increased. If this trend were to be applied to a real-life example, this would make a good deal of sense. An example would be for a particularly large retailer who sells bikes and wants to predict what to buy in the future. At a threshold of 75, not many demand numbers may fall below that value. Therefore, the retailer may be working with severely truncated data. Therefore, purchase quantities in the future may be too skewed towards the lower demand observations, in turn risking stock-outs. If, however, the retailer were to increase the threshold value of 150, i.e. holding starting with 150 bikes per day, there is a higher potential for there to be bikes left over and for the demand to be clearly seen. In essence, a higher threshold value can mean less truncated data and more information to make better decisions about forecasts of future inventory. Although it would be ideal to stock large quantities of inventory and in return have lots of information to make purchasing decisions in the future, it is not possible because of costs for inventory and obsolete items. Instead, retailers could use the ordered estimate approach to more accurately predict future demands.

Monetary Implications for the Seed Industry

The Matlab results were certainly promising for the ordered estimates approach, but there is another question regarding how much this new method could actually save for a company. To apply this thesis to a business case, we revisited the commercial seed producer's problem. We considered that the seed producer wanted to optimize its expected profit using the estimate of the yield distribution using the data available. Mathematically, the firm's problem is stated as:

$$\max_{\mathrm{Q}} -cQ + pE[\min(Q\,\tilde{y}, D)]$$

where *c* is the per acre production cost, *p* is the selling price per bag, *Q* is the number of acres used, $\tilde{y}$ is random yield, and *D* is the fixed demand. On further simplification the problem is written as:

$$\max_{\mathrm{Q}} \quad p\int_{y=0}^{y<\frac{D}{Q}} Qy * \phi\,(y)dy + P\int_{y>D/Q}^{\infty} D\phi(y)dy - CQ \qquad (5)$$

where $\phi(y)dy$ is the pdf (probability density function) of the yield. After completing this integral, the formula then turns into the following:

$$\max_{Q} pQ[-\sigma^2\phi(D/Q) + \mu\Phi(D/Q)] + PD[1 - \Phi(D/Q)] - cQ \qquad (6)$$

where $\phi\left(\frac{D}{Q}\right)$ is the pdf, while $\Phi(\frac{D}{Q})$ is the cdf, cumulative probability function. We set up this formulation in Solver in Excel with the price as 150, cost as 7500, 100 values of estimated means and standard deviations that come from the Matlab results for both approaches where the threshold value was 100, target mean was 100, and standard deviation was 30, demand as 10000, an actual mean of 100, and an actual standard deviation of 30. For each of the 100 instances, the solver optimized the profit function to provide the optimal acreage and the expected profit.

We ran the solver the model a total of 200 times, 100 times using the mean and standard deviation obtained using the naïve approach, and then 100 times using the mean and standard deviation using the order statistics approach. We then compared the difference in the expected profit from the two approaches. This analysis showed that on average the maximum profit under the old approach was $528, 529, while the new approach had a maximum of $572, 845. Therefore, using the order statistics based approach increases the expected profit by $44, 316, which is equivalent to 8% of profit.

This analysis reinforces that the ordered estimates approach is the preferred technique.

Chapter 6

Conclusion

The goal of this thesis was to test two separate approaches for estimating yields in the seed industry when truncated data was available. In the first approach only the available data were used to estimate the mean and standard deviation. In the second approach a weight was applied to each known observation using the sample size of all data. To test these two methods, Matlab code was developed to run a simulation based numerical study. The results revealed that the second approach is superior. The estimates of the mean and standard deviation obtained using this approach were consistently more accurate, on average, than the first approach. This trend persisted for varying values of standard deviation and truncation.

Finally, we applied these findings in the context of Dow AgroSciences. In test case with realistic data we found that the new approach could improve the expected profit for the firm by 8%. The future research should strive to implement these findings in practice.

Appendix

Matlab Code:

```
clc, clear

rng(1);

iterationssize=100000;

size=30;

HW1 = zeros(iterationssize,size);

trans = zeros(size,2);

for m = 1:iterationssize

   for n = 1:size

      HW1(m,n) = normrnd(0,1);

   end

end


HW1 = sort(HW1,2); % sorts rows in ascending order


Cv = cov(HW1);

alphafor10=mean(HW1,1);


for ex = 1:1000;

HW2 = zeros(1,size);


   for d = 1:size;

      HW2(1,d) = normrnd(100,20);

   end
```

```
HW2 = sort(HW2,2);

c=0;

for d = 1:size;

    if(HW2(1,d)<100)

      c=c+1;

    end

end


clear onesv;

Covar = Cv(1:c,1:c);

alpha=alphafor10(1:c);

ctemp=c;

for i=1:ctemp

  onesv(1:i)=1;

end


y=HW2(1:c);


temp = zeros(1, 2);

temp(:,1) = mean(y);

temp(:,2) = std(y);

thetasnaive(ex,:) = [temp(:,1)  temp(:,2)];


Sigma=inv(Covar);
```

```
p=[onesv'  alpha'];

pdash=transp(p);

theta=(inv(pdash*Sigma*p))*pdash*Sigma*y';

thetasorder(ex,:) = transpose(theta);

end



mean(thetasnaive)

mean(thetasorder)
```

Bibliography

Brida, J. G., Pereyra, J. S., & Scuderi, R. (2014). Repeat tourism in uruguay: Modelling truncated distributions of count data. *Quality and Quantity, 48*(1), 475-491.

Lloyd, E. H. (1952). Least-Squares Estimation of Location and Scale Parameters Using Order Statistics [Abstract]. *Biometrika,* Vol. 39, No. ½ (Apr., 1952), pp. 88-95.

Lowe, T. J., & Preckel, P. V. (2004). Decision technologies for agribusiness problems: A brief review of selected literature and a call for research.*Manufacturing & Service Operations Management*, *6*(3), 201-208.

Lu, X., Song, J.-S., & Zhu, K.. (2008). Analysis of Perishable-Inventory Systems with Censored Demand Data. *Operations Research*, *56*(4), 1034–1038.

ACADEMIC VITA

Academic Vita of Erik Zalewski
ejz5036@psu.edu

Education:

The Pennsylvania State University
Smeal College of Business
Class of 2016
Bachelor of Science Degree in Supply Chain and Information Systems
Minors in Information Systems Management and French

Thesis Title: Estimating Yields With Truncated Data
Thesis Supervisor: Saurabh Bansal

Internship Experience:

Ford Motor Company
May 2015 – August 2015
Materials Planning and Logistics Intern

State College Spikes
June 2014 – September 2014
Promotions and Community Relations Intern

Logan Capital Management
May 2013 – August 2013
Intern

Activities:

President – Clown Nose Club
September 2015 – May 2015

Webmaster – French Club
September 2015 – May 2015

Secretary – International Business Association
January 2015 – December 2015