DETERMINING THE APPLICABILITY OF THE TRIANGULAR DISTRIBUTION IN DESCRIBING AGRICULTURAL PROCESS ENERGY AND MASS FLOWS

By

LUIS R SERRANO

Bachelor of Science in Engineering in Mechatronics

Tecnológico de Monterrey Campus Chihuahua

Chihuahua, México

2009

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2011

DETERMINING THE APPLICABILITY OF THE TRIANGULAR DISTRIBUTION IN DESCRIBING AGRICULTURAL PROCESS ENERGY AND MASS FLOWS

Thesis Approved:

Dr. Scott Frazier

Thesis Adviser

Dr. Carol Jones

Dr. Timothy Bowser

Dr. Mark E. Payton

Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to acknowledge several individuals that helped me with suggestions, tips and comments that oriented me throughout this process. Without these individuals, this thesis would have not been completed. The order in which these individuals appear is not meant to emphasize their importance in this project.

I would like to start by thanking my advisor Dr. Scott Frazier for his tremendous passion that we both share on this field. Also, I would like to thank Dr. Carol Jones, a member of my committee, for being very supportive and willing to help whenever I came up with a question. Dr. Timothy Bowser, also a member of my committee, for his time and the advice he would give me in regards to my thesis.

Dr. Harry Field, because without him, I would not be sitting here writing his thesis. I would like to thank him for convincing me to do my Masters in Biosystems and Agricultural Engineering. When I first started commenting about going to graduate college, he directed me to the right department and the right people.

Dr. Glenn Brown, for the time we spent in his office discussing my future as a graduate student and for giving me the opportunity of doing my Masters in the Biosystems Department. Dr. Paul Weckler and Dr. Mark Wilkins, who reminded me of the basics of converting units and energy efficiencies. Dr. Ajay Kumar, for lending me his time to answer some questions related to this thesis. Dr. Gopal Kakani, and Dr. Yanqui Wu from the Plant and Soil Science Department, for giving me the opportunity to discuss with them basic concepts related with feedstock production. Dr. Carla Goad for answering all my questions and clearing all my doubts that deal with statistics. I would also like to thank Dr. William Raun, who would always make time in his tight schedule to discuss my doubts and problems. My friend Karthikeyan, for helping with my statistic principles and teaching me how to use SAS. My friend Celso Tamele for sharing his knowledge, advise and comments in regards to my thesis.

The entire staff of the Writing Center, who helped me with my writing abilities and corrected my grammar problems. My friends and colleagues from Oklahoma State University, who would stay up until late night in room 209 working on our projects.

I would also like to thank all the staff and faculty from Oklahoma State University. Melissa Moore and Nancy Rogers, for making my two years as a graduate student a very pleasant experience. Finally, I would like to thank Katy Kenslow and my family, who always walked with me throughout this whole process and made possible.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The beginnings of this research can be traced back when the author encountered an issue while working on a biofuel Industrial Ecology (IE) project. Industrial Ecology studies processes from a cradle to grave perspective, analyzing the environmental impact, energy and mass flows involved in a process. One of the main goals of IE is to change a process from a linear system into a closed loop system. Another goal of IE is to assimilate a concept that is able to minimize waste production while maintaining the process's productivity level. However, to achieve this goal, a significant amount of time, money and work has to be invested (Baas 2007).

While working on the biofuel IE project , the author used a Life Cycle Assesment (LCA) tool to map the energy and mass units that were going in and out of a given process (Figure 1). This LCA tool has been used in previous research (Lund, 2008) to analyze the energy, waste, and materials involved in the ethanol process. In this study, the author was trying to describe processes related with biofuel production from feedstock. However, after a short meeting with Dr. Scott Frazier from the Biosystems Department at Oklahoma State University on January 10th, 2011, the author was advised that the LCA tool was inefficient and time consuming due to its lack of boundaries when tracing energy and material input amounts backwards. This advice was also confirmed by author Biswas (2008) and the International Energy Agency (2000), suggesting in their articles that the LCA tool needs to be limited in order to produce reliable and accurate results.

As a possible solution, the author decided to build a scoping methodology that would allow researchers to forego the mapping process. This scoping methodology was going to be based on a signal to noise ratio that would compare the accumulated vartiations in a stream with the variation of the main output stream. This could allow researchers to save time without losing the efficiency of the process. Also, by using the variance accumulated in a backtracking analysis and with the aid of statistical methods, limits could be drawn on the mapping of inputs. However, the data collected by the author was composed primarily of point estimates that lacked variation. Also, the author realized that there was not enough information available that would describe the distribution of the biofuel production processes. Due to this lack of information, the author was unable to assume a probability distribution that could describe the behavior of the energy and mass estimates described in Figure 1.

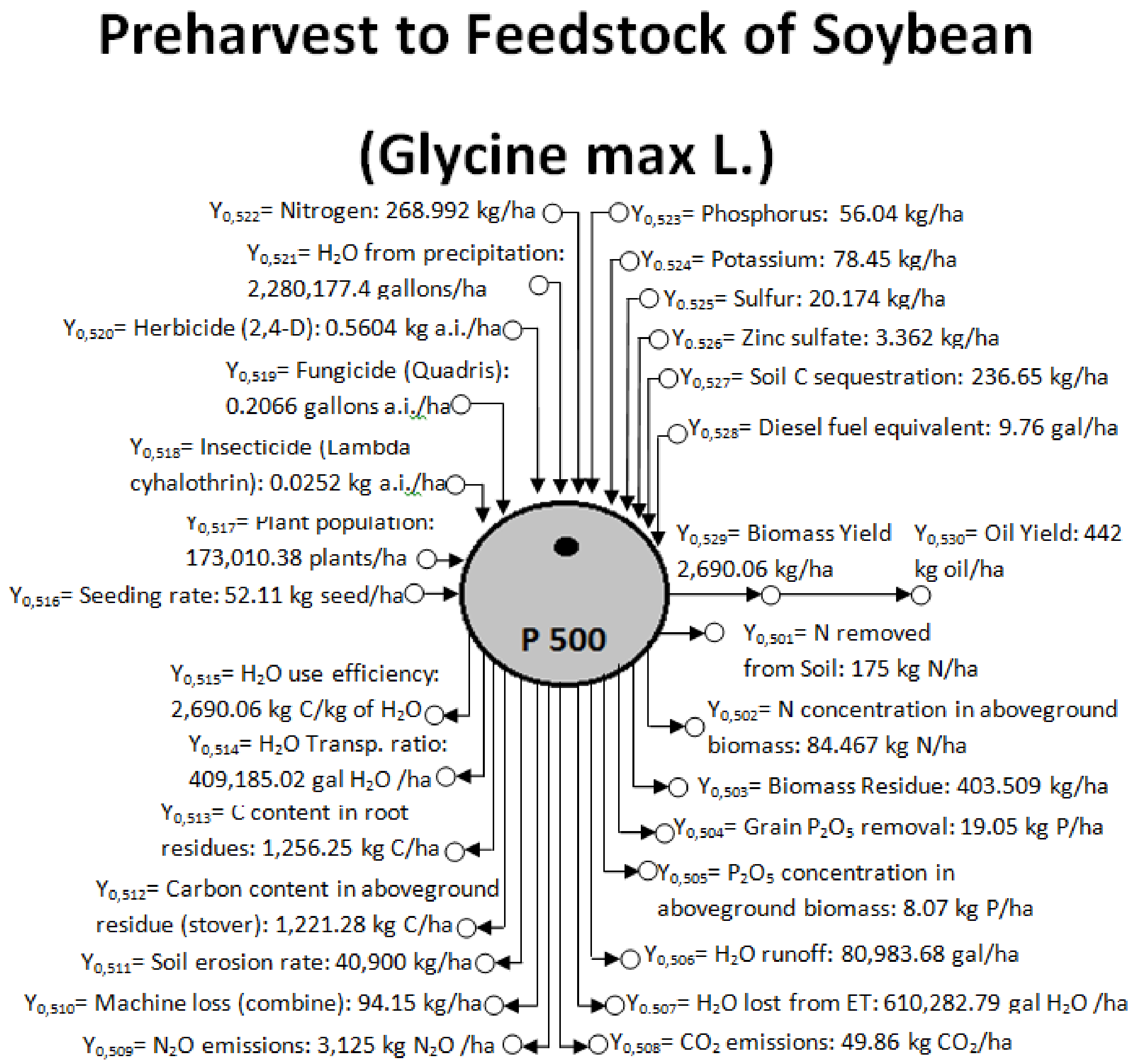

Preharvest to Feedstock of Soybean
(Glycine max L.)
$Y_{0,522}$= Nitrogen: 268.992 kg/ha
$Y_{0,523}$= Phosphorus: 56.04 kg/ha
$Y_{0,521}$= H_2O from precipitation: 2,280,177.4 gallons/ha
$Y_{0.524}$= Potassium: 78.45 kg/ha
$Y_{0.525}$= Sulfur: 20.174 kg/ha
$Y_{0,520}$= Herbicide (2,4-D): 0.5604 kg a.i./ha
$Y_{0.526}$= Zinc sulfate: 3.362 kg/ha
$Y_{0,519}$= Fungicide (Quadris): 0.2066 gallons a.i./ha
$Y_{0,527}$= Soil C sequestration: 236.65 kg/ha
$Y_{0,528}$= Diesel fuel equivalent: 9.76 gal/ha
$Y_{0,518}$= Insecticide (Lambda cyhalothrin): 0.0252 kg a.i./ha
$Y_{0,517}$= Plant population: 173,010.38 plants/ha
$Y_{0,529}$= Biomass Yield 2,690.06 kg/ha
$Y_{0,530}$= Oil Yield: 442 kg oil/ha
$Y_{0,516}$= Seeding rate: 52.11 kg seed/ha
P 500
$Y_{0,501}$= N removed from Soil: 175 kg N/ha
$Y_{0,515}$= H_2O use efficiency: 2,690.06 kg C/kg of H_2O
$Y_{0,502}$= N concentration in aboveground biomass: 84.467 kg N/ha
$Y_{0,514}$= H_2O Transp. ratio: 409,185.02 gal H_2O /ha
$Y_{0,503}$= Biomass Residue: 403.509 kg/ha
$Y_{0,513}$= C content in root residues: 1,256.25 kg C/ha
$Y_{0,504}$= Grain P_2O_5 removal: 19.05 kg P/ha
$Y_{0,505}$= P_2O_5 concentration in aboveground biomass: 8.07 kg P/ha
$Y_{0,512}$= Carbon content in aboveground residue (stover): 1,221.28 kg C/ha
$Y_{0,511}$= Soil erosion rate: 40,900 kg/ha
$Y_{0,506}$= H_2O runoff: 80,983.68 gal/ha
$Y_{0,510}$= Machine loss (combine): 94.15 kg/ha
$Y_{0.507}$= H_2O lost from ET: 610,282.79 gal H_2O /ha
$Y_{0,509}$= N_2O emissions: 3,125 kg N_2O /ha
$Y_{0,508}$= CO_2 emissions: 49.86 kg CO_2/ha

Figure 1. Input/output diagram from an Industrial Ecology project analyzing soybean production

Two possible statistical solutions that could solve the issues described previously would be the Beta distribution and the Triangular distribution. Both these distributions are used in management tools like the Project Evaluation and Review Technique (PERT) and Critical Path Method (CPM) to estimate project completion times based on maximum and minimum values. However, the Beta distributions requires of a considerable amount of data to describe a population's distribution (Zou and Normand 2001). Not being able to gather enough information, the author decided to use a Triangular distribution instead to describe the variance and find the estimates of the energy and mass inputs.

Unlike the Beta distribution, the Triangular distribution only requires one experienced individual to obtain an estimate and calculate a variance. This variance is based on the experience of operators or experts related with the task being analyzed (Vose 2008). Finally, the author decided to perform a statistical test to determine the applicability of the triangular distribution in agricultural processes where input and output data can be scarce. This statistical test was based on a hypothesis that there is no statistical difference between the random variables obtained from a triangular distribution and those obtained from the Census of Agriculture (USDA 2008) when applied in agricultural processes. To perform this test, the author decided to analyze the soybean yields in the state of Oklahoma. This analysis involved a quantitative study that gathered information from a questionnaire.

The results obtained from these questionnaires will help to reject or evaluate the null hypothesis, and also determine the applicability of the triangular distribution in describing agriculture process energy and mass flows. Also, if the author is able to prove that this hypothesis is true, meaning that the researcher will not be rejecting this hypothesis, other areas besides agriculture could benefit from using this triangular distribution. Some of these other areas that could implement this statistical tool would be Industrial Ecology. By capturing three variables from the process being analyzed, based on an expert’s opinion, a researcher could build a triangular distribution and obtain the variance required by scoping methodology tools used in Industrial Ecology projects.

1.1 OBJECTIVES

Now that the purpose of this study has been described, the objectives are listed below:

1. Decrease the amount of time and money invested in collecting data by using a triangular distribution instead to represent a statistical behavior of a population.
2. Collect a certain amount of data that allows a researcher to be 95% confident that the triangular distribution represents the population being analyzed.
3. Determine the applicability of the triangular distribution in agricultural process energy and mass flows by comparing the UDSA database of soybean yields in the state of Oklahoma with the triangular distribution obtained in this study.
4. Create a triangular distribution based on expert's opinions to represent a known or unknown distribution.
5. Establish if the triangular distribution could be utilized to describe the variance in energy and mass input streams found in Industrial Ecology projects.

CHAPTER 2

LITERATURE REVIEW

2.1 INDUSTRIAL ECOLOGY

Industrial Ecology is a study that not only evaluates the flow of materials or economics involved in a process, but also provides with tools that allow measuring the environmental impact of a system. Based on how the life cycle, ecosystem, and environment behave, one of the main goals of Industrial Ecology is to replace the linear process adopted by most industries with a closed loop model used by nature. To achieve this goal, Industrial Ecology could use statistical tools to produce accurate estimates, allowing researchers to describe mass and energy flows along process streams. One of the tools currently used in I.E. to track energy and mass flows involved in processes from a cradle-to-grave perspective is the Life Cycle Assessment (LCA) tool. This tool besides analyzing the behavior of industrial processes, it also describes agriculture techniques (Korhonen 2004; Seager and Theis 2002; Giurco, Cohen et al.). The following section will describe the advantages and disadvantages of this LCA tool.

2.2 LIFE CYCLE ASSESSMENT (LCA) TOOL

The Life Cycle Assessment (LCA) tool is mainly used in I.E. projects to describe the inputs and outputs that go in and out of processes. The LCA tool uses a cradle-to-grave analysis, where all the energy and mass units are described from the beginning until the end of a process.

However, the LCA tool is considered to be time-consuming and expensive (Kapur and Graedel 2002), mainly because it lacks of boundaries that indicate when to stop mapping the variables involved in a system. According to Azapagic (1999), the LCA tool may be considered an option when trying to solve environmental problems; but when dealing with process efficiencies, it can still be somewhat inappropriate because it tries to consider all the factors involved in a process. A solution that could be implemented to solve this issue would be to use a signal to noise ratio to compare the accumulated variation of input stream statistics with variation of the main output stream. This scoping methodology could be implemented by researchers to cease mapping of energy streams when the accumulated variation in the estimated values exceeds the variation of the main output stream. This scoping methodology would help the LCA tool to establish boundaries and increase the efficiency of mapping processes. The next section will describe the scoping methodology that the author chose to analyze in this study.

2.3 ESTABLISHING A SCOPING METHODOLOGY

Regardless of the overall analysis that Industrial Ecology accomplishes by using modeling tools like the LCA to describe the economic, social and environmental impact of a system; the LCA tool is considered to be a never ending process that lacks a scoping methodology (Gao 2006). The scoping methodology used by the author is based on a statistical method that analyzes the accumulated variance of estimated variables in an energy stream. When a researcher analyzes the estimated variables in an energy stream, the variation of such variables should also be taken into account. This variation can be used to build a signal to noise ratio described by a standard error according to the following equation (Brown 1999),

$$SE = \frac{s}{\sqrt{n}}$$

Where (SE) represents the standard error, (s) describes the standard deviation of a sample and (n) represents the sample size of an experiment. The accumulated variation will allow the researcher to calculate a standard deviation, which will then be used to obtain a standard error. The standard error (SE) of the main output shall then be compared with the accumulated standard error. According to the author, when the accumulated SE exceeds the SE that corresponds to the main output of the system, the mapping process should be stopped. However, to use this SE, a researcher must have statistics with variation information for the various input streams. For this study, such variation information was not available. To solve this issue, the author decided to find a probability distribution that would describe the variance of the data being analyzed. However, before choosing a statistical solution, the author will first explain the main differences between a parametric and a non-parametric distribution to analyze the options that would be available to establish a scoping methodology.

2.4 DIFFERENCE BETWEEN PARAMETRIC AND NON-PARAMETRIC DISTRIBUTION

Parametric distributions describe the behavior of a researcher's collected information according to a probability distribution. This means a researcher assumes the collected information will behave according to a specific parametric distribution, which might be normal, uniform, exponential, etc. A parametric distribution is mainly described by a mean value (μ) and a standard deviation (σ). These types of distributions are considered to have a higher statistical power when compared to non-parametric distributions. A statistical power refers to the probability of not rejecting a null hypothesis H_0 (failing to reject) when a researcher should indeed reject the hypothesis. This scenario described previously is also known as a Type II error. A distribution with a higher statistical power will result in a lower probability of making a Type II error.

The higher statistical power may be due to the typically larger sample size or because there is a higher significance level (α) involved in the parametric distributions. However, a parametric distribution is also known for being less robust. If a distribution is less robust, it is considered to have a higher probability of being affected by outliers. Also, parametric distributions are known for using larger sample sizes if compared with non-parametric distributions. Examples of statistical tests used in parametric distributions are the Fisher's F tests, the Student's t test, and Chi-square test.

However, there are also other types of distributions classified as non-parametric.. Nonparametric distributions do not assume that the data obtained by a researcher follows a specific probability distribution. They are used when there is not enough information to represent a population's distribution (Johnson 1997; Mohammadi et al. 2007). Unlike the statistical tests utilized in the parametric distributions, the non-parametric distributions use ranking methods. To differentiate between a parametric and a nonparametric distribution, the author started by plotting the data obtained during this study in a histogram to visually determine if the information was represented by a parametric or a non-parametric distribution. After analyzing the frequency distribution, the author used the statistical program Arena ® 12 to analyze the data and fit a distribution accordingly. Once a researcher knows what type of distribution represents the data being analyzed, statistical tools can be chosen to test a hypothesis. The next section will describe the Normal distribution, followed by the Beta distribution.

2.5 NORMAL DISTRIBUTION

Considered one of the most common parametric distribution functions when using independent variables (Joyce 2006), the normal distribution (also called the Gaussian distribution) represents a number of (Y) random variables that can take any value from ($-\infty \ to \ \infty$). Often referred to as the bell curve, this distribution is described by having a variance (σ =1) and being symmetric about the mean value (μ=0). Using the probability tables (Z) of the normal distribution, a researcher can determine the probability of occurrence of a certain variable or the confidence intervals at which a variable is expected to occur. The following equation describes this probability (Freund and Wilson 2003):

$$Z = \frac{(Y - \mu)}{\sigma}$$

Where Y represents the value being tested, (μ) represents the population mean and (σ) represents the standard deviation of the population. However, in order to use the previous equation, the sample size (n) must be equal to or greater than 30. Also, the probability of every random variable falling under the curve, when added all together, must be equal to 1. The confidence intervals can be defined by the researcher according to the standards of deviation and the quality standards set by the customer. This distribution is commonly used by researcheres when there is enough information to assume the data behaves according to a normal distribution. However, there was not enough information collected in this study to be able to make this assumption, forcing the author to keep looking for other options. The following section will describe the Beta distribution.

2.6 BETA DISTRIBUTION

Used along with the triangular distribution as a decision management tool, the Beta distribution approximates the mean and variance used in the PERT and CPM tools. The following equations describe how the mean and variance are calculated (Vose 2000):

$$\mu\,(x) = \frac{A + 4M + B}{6}$$

$$\sigma(x) = \frac{B - A}{6}$$

Where (μ) is the expected value, (σ) is the variance of a variable compared to its mean, and the variables (A), (M) and (B) have the same meaning as in the triangular distribution. The Beta distribution uses a maximum and minimum value like the triangular distribution to delimit project completion times. However, according to Keefer and Verdini (1993), the Beta distribution is considered to be a poor estimation tool because it calculates the mean variance with a 40% and 549% error respectively. Nonetheless, Keefer and Bodily (1983) consider this Beta distribution to be a better approximation than the one used in PERT analysis. However, if an accurate triangular distribution is built based upon a quantitative study, the error percentage could be less than the one estimated by Keefer (1993).This could add more credibility to the three-point estimation tools like the triangular distribution and might help to reduce the time and money spent by researchers on more complex statistical tools. Also, the Beta distributions requires a considerable amount of information to describe a population's distribution (Zou and Normand 2001). Due to the lack of information in this study, the researcher chose the Triangular distribution to determine its applicability in describing agricultural process energy and mass flows. The author will describe the triangular distribution in following section.

2.7 TRIANGULAR DISTRIBUTION

First used in the year 1755 by the Englishman Thomas Simpson, the triangular distribution is considered to be the first continuous distribution model to appear in the 18th century (Kotz and Rene van Dorp 2004). According to Seal (1949), Simpson (1757) was trying to mathematically represent the error experienced by astronomers when they would use their instruments or take their measurements based on the human eye. Simpson (1757) assumed that these astronomers would calculate their averages based on a lower and upper limit, creating a discrete assymetric triangular distribution (Kotz and Rene van Dorp 2004).

The following equation represents the probability of a variable (x) assuming a value between two established variables (A and B). These upper and lower variables, identified with the variables A and B, respectively, are the basis for the probability density function for continuous distributions described in the following equation (Sleeper 2006).

$$\int_{A}^{B} f(x)dx = P[A < X < B]$$

Where A and B represent real values.

Between these two estimates, there is a third variable called the mode. This mode, usually defined by the letter (M), is the value that repeats the most in a current observation. As described in Figure 2, the total area of the triangle can be divided by two separate sections A_1 and A_2. If the triangle is non-symmetric, these two areas will not be equal and the mode will not represent the midpoint of the triangle.

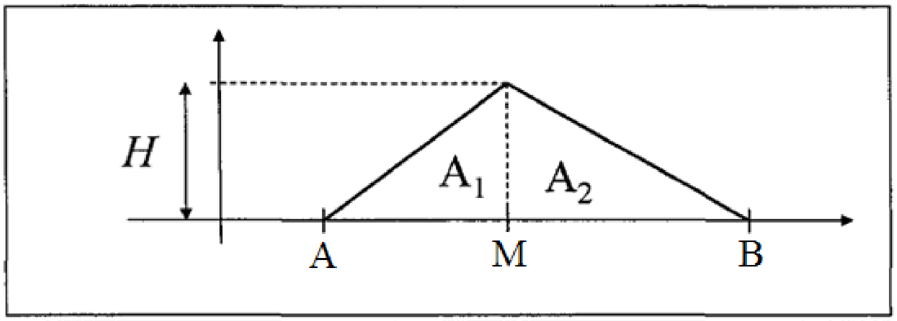


Figure 2. Non symmetrical triangular distribution. Source: (Kotz and Rene van Dorp 2004).

If the area of a triangle is:

$$Area = \frac{Base\ x\ Height}{2}$$

And according to Figure 2, $A_1 + A_2 = A_{TOTAL,}$ then the equation could be described as:

$$A_{Total} = \frac{(M-A)H}{2} + \frac{(B-M)H}{2} = 1$$

where the variable H could be substituted by:

$$H = \frac{2}{B-A}$$

In summary, the probability of a value (x) falling in between the values A and B can be described as (Kotz and Rene van Dorp 2004):

$$f\ (X|A, M, B) = \begin{cases} \frac{2}{B-A}\frac{X-A}{M-A}, for\ A \leq X \leq M, \\ \frac{2}{B-A}\frac{B-X}{B-M}, for\ M \leq X \leq B, \\ \quad 0, \qquad elsewhere \end{cases}$$

The mean value can be obtained by using the following equation:

$$Mean = \frac{A+B+M}{3}$$

The variance of a value in this type of distribution when compared to its mean is obtained through the following equation:

$$Variance = \frac{A^2 + B^2 + M^2 - AB - AM - BM}{18}$$

Finally, the next equation is used to calculate the standard uncertainty when using a triangular distribution is:

$$Standard\ Uncertainty = \frac{B-A}{2\sqrt{6}}$$

Due to its simplicity, basic equations, and by considering that the variables (A), (B) and (M) originate from the *opinion* of people, the triangular distribution has not been considered a reliable statistical representation (Weissberg and Buker 2005). However, researchers (Love and Goodman 2001; Administration 2002; Montville, Chen et al. 2002; Chen 2007) opt to use this statistical tool when the following criteria are met:

1. The lack of a known distribution.
2. Lack of time and money to work on a more elaborate statistical analysis.
3. The upper, lower and the most common outcomes are the only known variables.

Studies such as Industrial Ecology could also benefit from using this distribution. Considered a difficult task to collect the required information for a LCA project (Davis, Nikolic et al. 2010); the triangular distribution could help a researcher to collect such information more efficiently. Risk management tools such as the Critical Path Method (CPM) and the Project Evaluation and Review Technique (PERT) already use the triangular distribution to estimate project completion times.

Although the Beta distribution could be used to provide with estimates according to a low and a high value, there was not enough data available in this study that would allow the author to use such distribution. However, the triangular distribution only requires the opinion of one expert to obtain an estimate and calculate a variance. After analyzing the possibilities of establishing limits on a LCA tool, the author decided to use a triangular distribution to obtain estimates according to an expert's opinion related with soybean production in the state of Oklahoma. This triangular distribution could also be used to calculate the variance of the estimates obtained during this study and utilize that variance to implement a signal to noise ratio. This signal to noise ratio could then be used to describe the standard error of a variable.

The following chapter will describe how the author will use this triangular distribution to obtain estimates, calculates a variance, and at the same time test the applicability of the triangular distribution if it were to be used to describe agricultural and industrial processes.

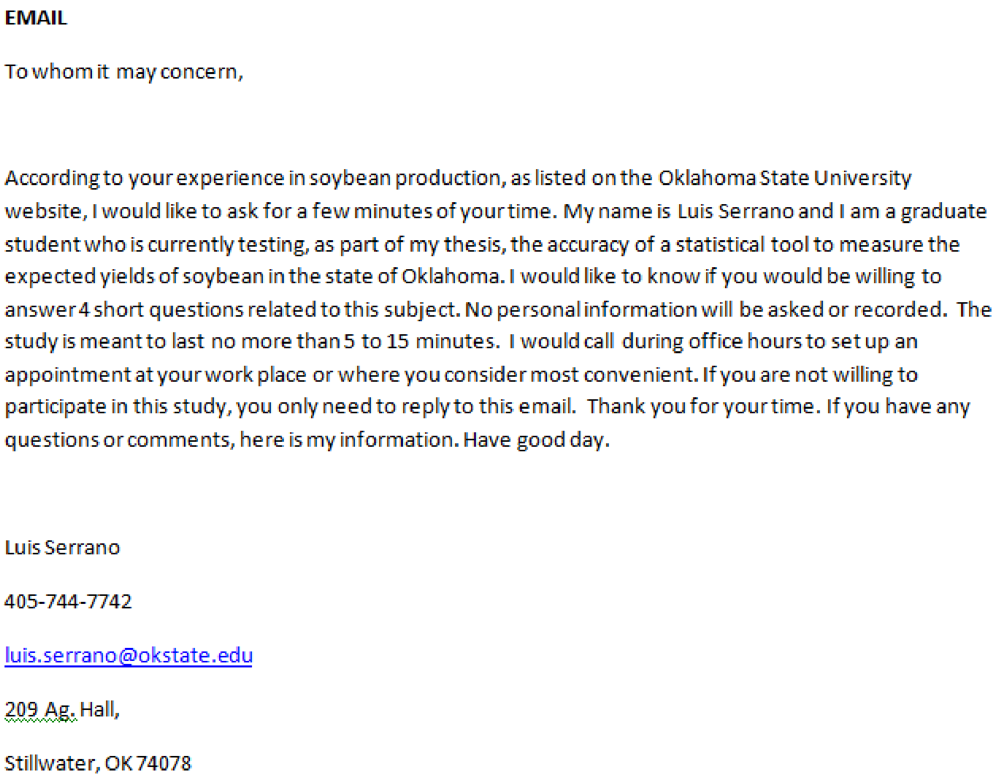
CHAPTER 3

METHODOLOGY

Soybean is considered to be one of the most important crops worldwide (Armstrong, Arnall et al. 2009). It is also considered to be a good candidate for producing biodiesel (Laboratory 2011). Oklahoma, being recognized for growing this crop, was chosen in this quantitative study to determine the applicability of the triangular distribution when estimating soybean yields. If the triangular distribution, with the help of statistical tests, could be proven not to be statistically different from the normal distribution, the time and money spent collecting data might be significantly reduced. The population of individuals used for this study is composed of three different groups. Every individual in every group is related with the production of soybean, providing credible answers based on their experience. However, this study did not include the opinion of students related to this crop since they might not share the same level of experience as the other individuals. If they were included in this study, the possibility of affecting our results and obtaining skewed estimates would be much greater.

Faculty members from the Oklahoma State University (OSU) from the Biosystems and Agricultural Engineering Department, the Agriculture and Economics Department and the Plant and Soil Sciences Department responded a questionnaire that took part in this study. This group of seven faculty members responded to four questions related to soybean production in the state of Oklahoma.

A second group of ten members from OSU, composed by Superintendents and Assistant Superintendents from experiment stations in Oklahoma, also provided their feedback on these questionnaires. Finally, a third group consisting of five farmers provided their feedback by answering the same questionnaire. This group was contacted based on recommendations made by faculty members and superintendents who answered this survey. Faculty members and superintendent's contact information was obtained through the OSU website. Figure 3 displays the email used to recruit the volunteers who participated in this study.



EMAIL

To whom it may concern,

According to your experience in soybean production, as listed on the Oklahoma State University website, I would like to ask for a few minutes of your time. My name is Luis Serrano and I am a graduate student who is currently testing, as part of my thesis, the accuracy of a statistical tool to measure the expected yields of soybean in the state of Oklahoma. I would like to know if you would be willing to answer 4 short questions related to this subject. No personal information will be asked or recorded. The study is meant to last no more than 5 to 15 minutes. I would call during office hours to set up an appointment at your work place or where you consider most convenient. If you are not willing to participate in this study, you only need to reply to this email. Thank you for your time. If you have any questions or comments, here is my information. Have good day.

Luis Serrano

405-744-7742

luis.serrano@okstate.edu

209 Ag. Hall,

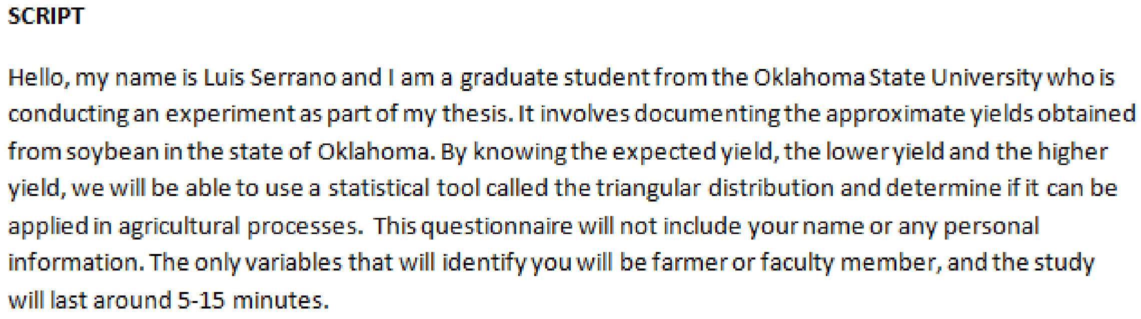
Stillwater, OK 74078

Figure 3.Email recruiting participants for a questionnaire used in this study.

After waiting one week and not receiving a reply, the next step was to call the participant during office hours to schedule an appointment. During this phone call, the interviewer would describe the purpose of this study and explain to the participant why they were selected.

The interviewer would also explain that their personal information was not being recorded for this study. This allowed participants to provide with more accurate estimates without affecting their reputation or results of this study. After giving this explanation, an appointment with a faculty member would be scheduled for a five to fifteen minute meeting at their workplaces during office hours.

When the day of the interview arrived, the script that appears in Figure 4 would be addressed to remind the interviewee about the purpose of this study.

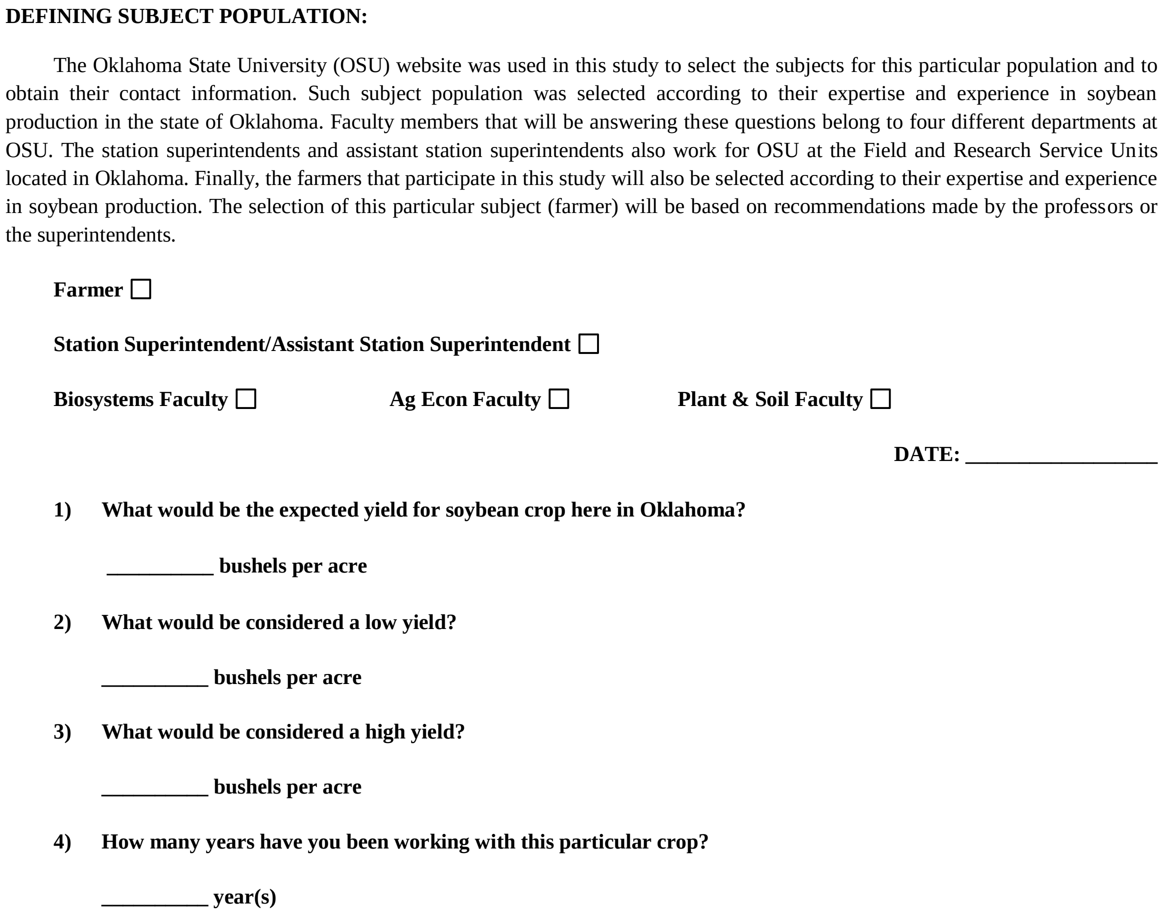


SCRIPT

Hello, my name is Luis Serrano and I am a graduate student from the Oklahoma State University who is conducting an experiment as part of my thesis. It involves documenting the approximate yields obtained from soybean in the state of Oklahoma. By knowing the expected yield, the lower yield and the higher yield, we will be able to use a statistical tool called the triangular distribution and determine if it can be applied in agricultural processes. This questionnaire will not include your name or any personal information. The only variables that will identify you will be farmer or faculty member, and the study will last around 5-15 minutes.

Figure 4. Script stated to the interviewee to describe the objective of this study

A questionnaire (Figure 5) was used to obtain data from a sampling population. The questionnaire used was not extended to the interviewee to allow a more personal interaction. According to (Dawson 2009), this questionnaire could be classified as close-ended, following a format where it only focuses on obtaining certain numbers, and does not look for any explanation of how the participants obtained these numbers. It can also be classified as a structured interview, mainly used for quantitative and not qualitative studies. A structured interview consists of a number of defined questions whose format remains unchanged throughout the sampling process.



DEFINING SUBJECT POPULATION:

The Oklahoma State University (OSU) website was used in this study to select the subjects for this particular population and to obtain their contact information. Such subject population was selected according to their expertise and experience in soybean production in the state of Oklahoma. Faculty members that will be answering these questions belong to four different departments at OSU. The station superintendents and assistant station superintendents also work for OSU at the Field and Research Service Units located in Oklahoma. Finally, the farmers that participate in this study will also be selected according to their expertise and experience in soybean production. The selection of this particular subject (farmer) will be based on recommendations made by the professors or the superintendents.

Farmer □

Station Superintendent/Assistant Station Superintendent □

Biosystems Faculty □ **Ag Econ Faculty** □ **Plant & Soil Faculty** □

DATE: __________________

1) **What would be the expected yield for soybean crop here in Oklahoma?**

__________ bushels per acre

2) **What would be considered a low yield?**

__________ bushels per acre

3) **What would be considered a high yield?**

__________ bushels per acre

4) **How many years have you been working with this particular crop?**

__________ year(s)

Figure 5. Questionnaire used in this study.

After answering this questionnaire, the faculty members and superintendents were asked if they knew of farmers related with this crop in the state of Oklahoma who would be willing to answer this questionnaire. If such information was given, the interviewer would contact the farmer to obtain the three variables used in the triangular distribution by using the questionnaire described previously.

The following diagram (Figure 6) summarizes the methodology followed to recruit the participants for this study.

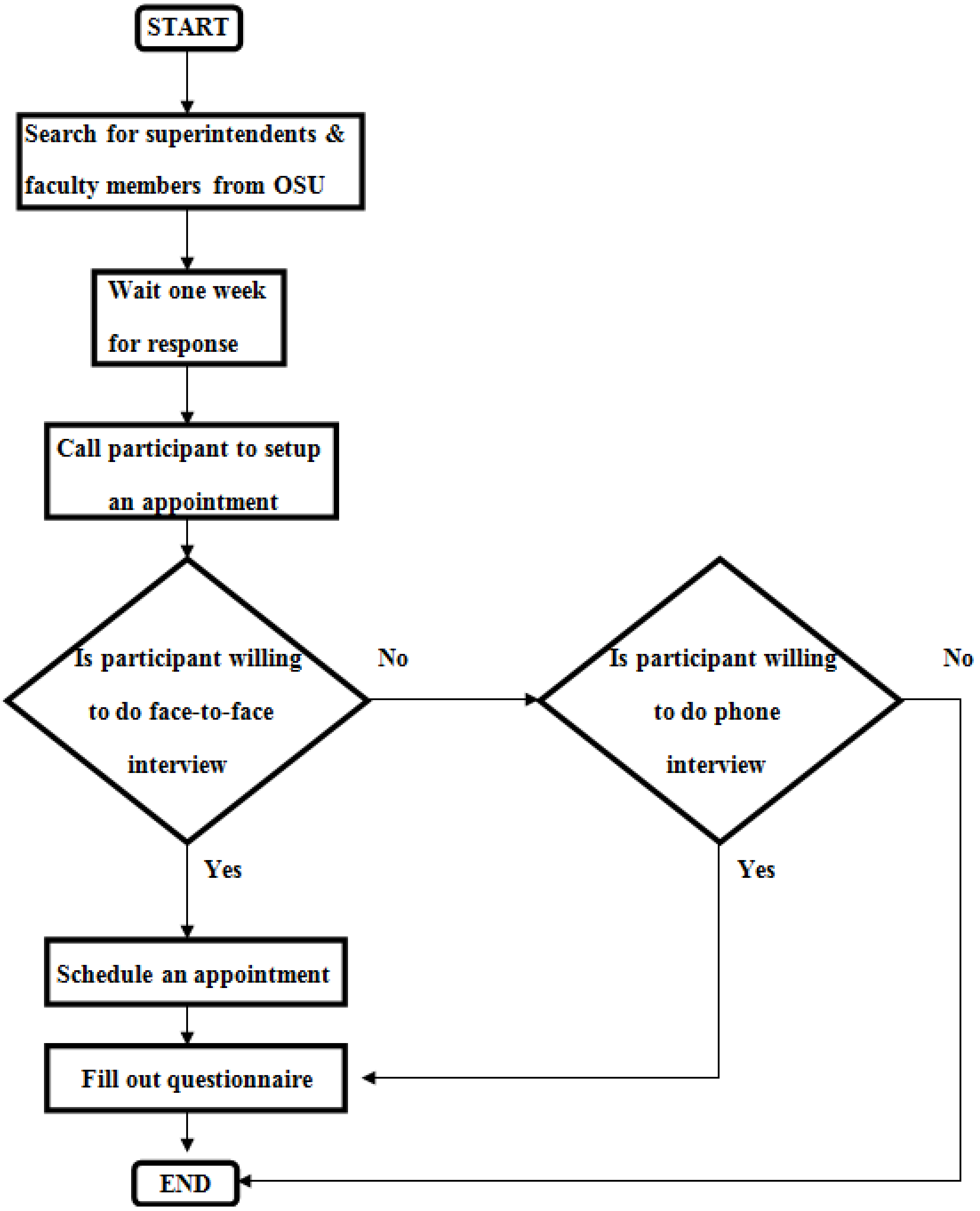


Figure 6. Process flow diagram describing participant recruiting for this study

Once all the information was gathered from the questionnaires, the (A) and (B) variables were used as a range for building the triangular distribution. Also, the expected outcome was used as a mean value. Once these three values were obtained, the mode variable (M) was calculated.

The next step was to find out what type of distribution this data followed. After selecting the appropriate distribution by using the Arena 12 ® software, statistical tools were chosen to test the hypothesis. These statistical tools compared the data obtained from the questionnaires with the soybean database available from the USDA web page. In this study, the author will refer to the soybean yield database from the years 1961 to 2010 as the “model data”. This model data was available through the USDA web page. Also, an additional database was captured containing the soybean yields from the year 2010 in the state of Oklahoma. The author will refer to this database as the “recent data” This additional database and the model data from the years 1961-2010 would be used in a Kruskal Wallis test. These two databases available from the USDA web page, along with sampled data, would be used in a Kruskal Wallis test to see if the data originates from a single population.

The following diagram (Figure 7) summarizes the statistical tests used to compare the sampled data with the soybean yield according to the model data and recent data obtained from the USDA web page. Once the appropriate statistical tools and the corresponding statistical tests have been made, a second statistical test will take place.

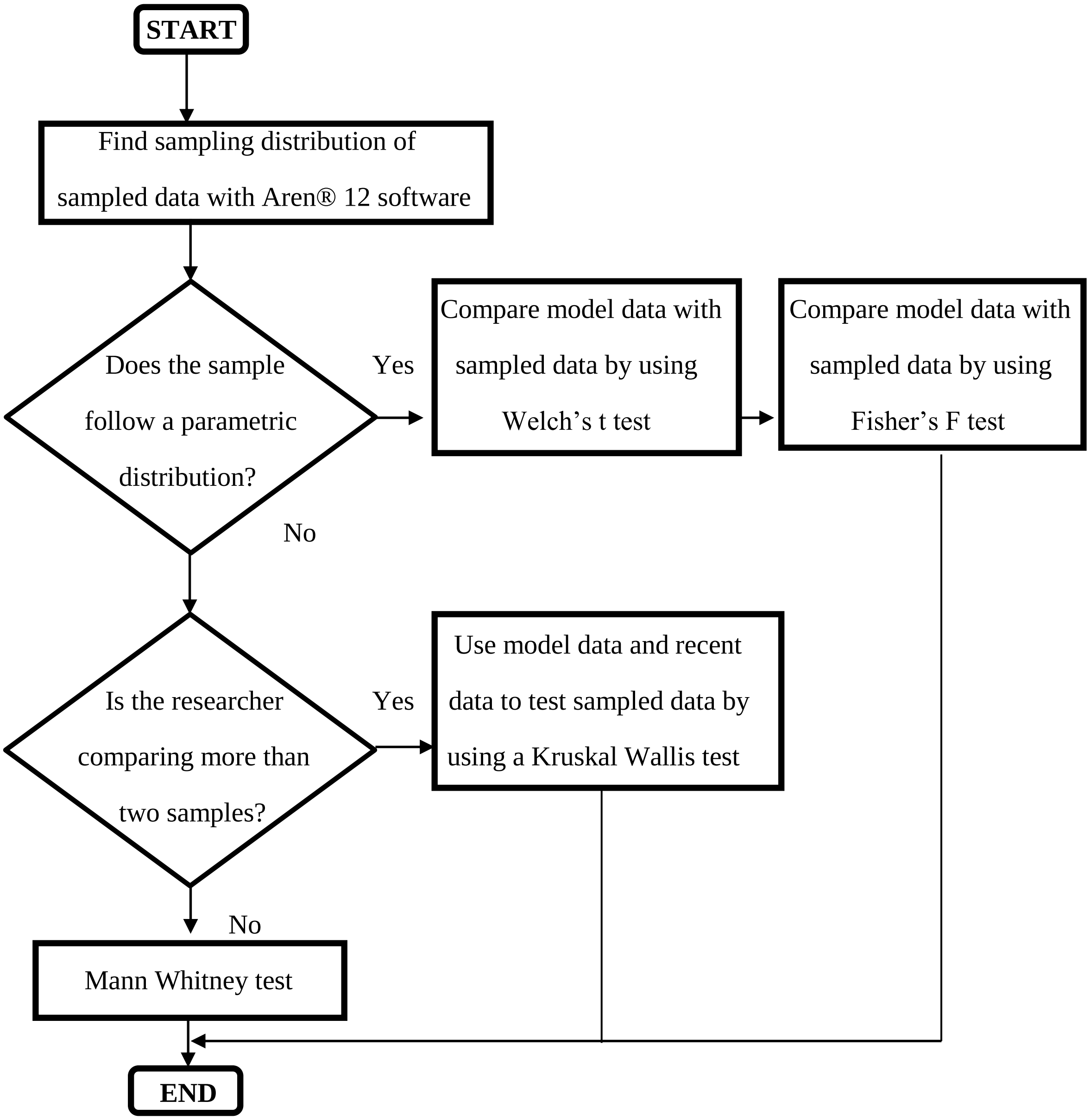

START
Find sampling distribution of sampled data with Aren® 12 software
Does the sample follow a parametric distribution?
Yes
Compare model data with sampled data by using Welch's t test
Compare model data with sampled data by using Fisher's F test
No
Is the researcher comparing more than two samples?
Yes
Use model data and recent data to test sampled data by using a Kruskal Wallis test
No
Mann Whitney test
END

Figure 7. Diagram describing statistical tests performed in this study according to sample's distribution.

This second statistical test, described in Figure 8 will compare the individual sampling groups (farmers, faculty members and superintendents) that composed the overall population with the model data. This test allowed the author to analyze each group individually and determine their influence on the results obtained from the first statistical test.

The distribution of each group was modeled with the help of the Arena 12 ® software. Once the distributions were identified, appropriate statistical tools were selected. These statistical tools allowed the comparison of each sampling group with the model data to determine if there was a statistical difference or not.

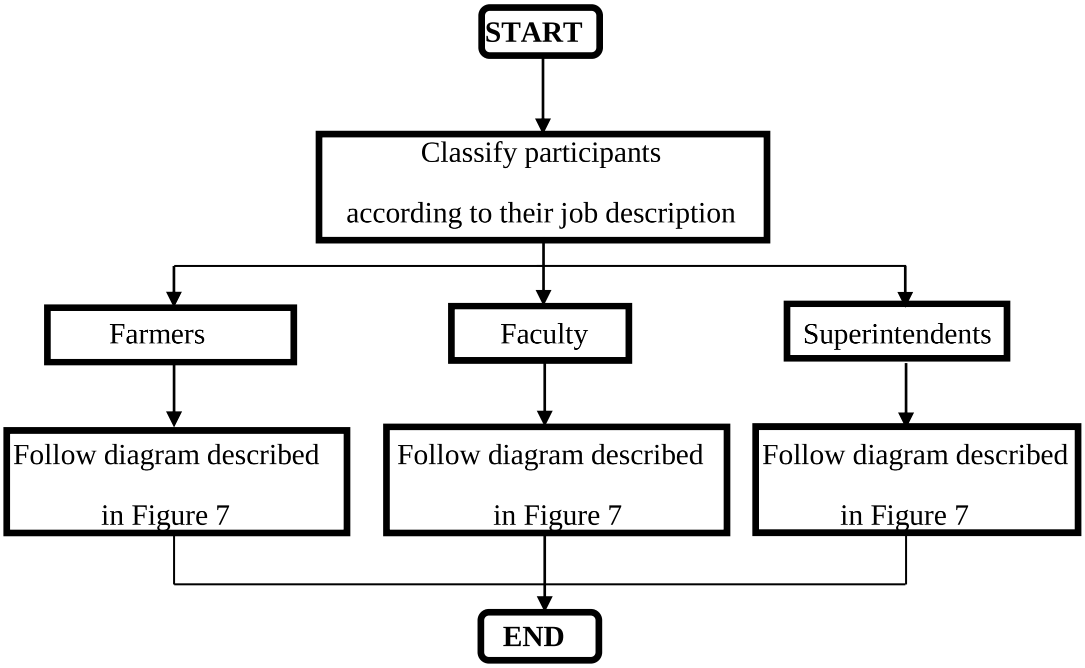

START
Classify participants according to their job description
Farmers
Faculty
Superintendents
Follow diagram described in Figure 7
Follow diagram described in Figure 7
Follow diagram described in Figure 7
END

Figure 8. Diagram describing the second statistical test where groups were analyzed individually

After defining if the sampling groups were or were not statistically different from the model data, a third statistical test took place. This statistical test stratified the sampling population according to the experience level of each individual related with soybean production. The process of stratification consists of separating the sampling population into mutually exclusive and collectively exhaustive subgroups. A mutually exclusive event assigns only one subgroup per participant, which means that a participant cannot belong to more than one subgroup. A collectively exhaustive event classifies every member of a group into a smaller group without leaving any participant unclassified. This classification is described in the following table (Table 1):

Table 1. Groups classified according to their experience level

EXPERIENCE LEVEL	YEARS OF EXPERIENCE
Not so experienced	$0 \leq x \leq 5$
Experienced	$5 \leq x \leq 10$
Expert	$x > 10$

Irwin (2006) suggests that separating the sampling group into smaller subgroups with a common trend might reduce the variance of such analysis if compared to an overall analysis of the variance of a whole population. These subgroups also allowed to test and analyze the influence and effect of the experience level in a triangular distribution. After classifying every individual according to the experience level and determining the distribution, statistical tests were chosen to compare these three groups with the model data. Finally, after analyzing all of the results obtained from the three statistical tests, the author was able to determine if there is a statistical difference between the triangular distribution and the normal distribution. Figure 9 summarizes the third statistical test described previously.

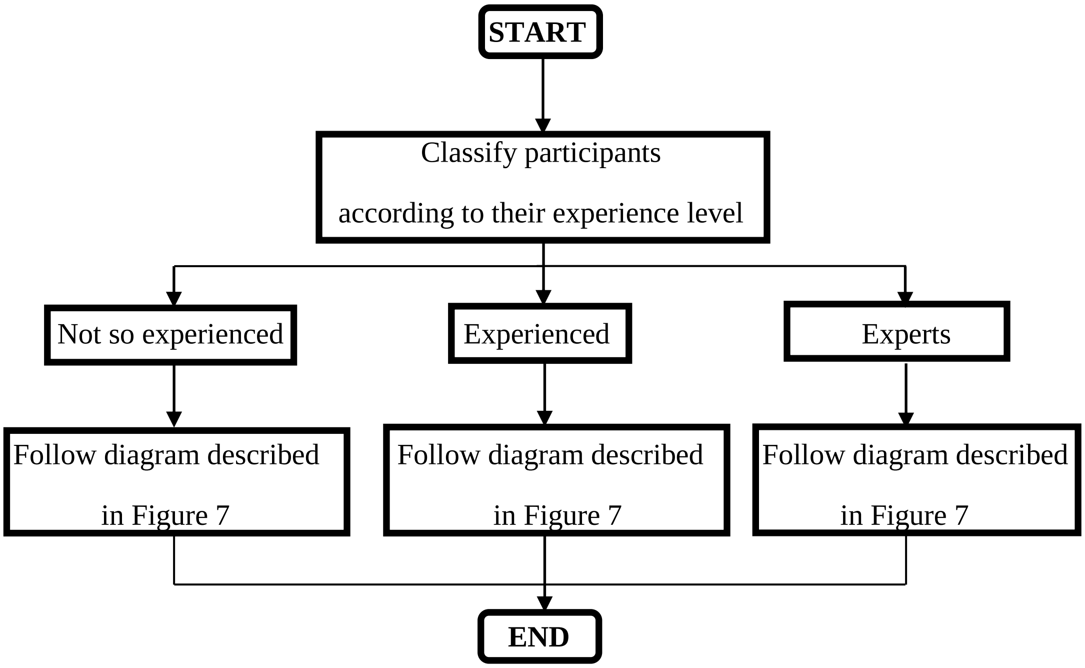

START
Classify participants according to their experience level
Not so experienced
Experienced
Experts
Follow diagram described in Figure 7
Follow diagram described in Figure 7
Follow diagram described in Figure 7
END

Figure 9. Diagram describing third statistical test.

3.1 SNOWBALL SAMPLING

A snowball sampling tool was used to obtain the participants for this study. This sampling tool is a nonprobability method that selects sampling individuals in a nonrandom matter according to a certain criteria (Goodman 1961). The sampling individuals then recommend a number of participants who fit the same criteria and the sampling process starts again. The process may stop whenever the researcher has determined that there is enough information collected or when there is no more individuals being recommended. This nonrandom sampling tool is considered an option when there is a restricted budget to hire a workforce and the sampled population is difficult to contact (Castillo 2009). For this study, the author was looking for participants related with soybean production in the state of Oklahoma. According to the author, the information obtained from the questionnaires should be based on the participant’s experience without having to consult any articles, books or literature.

The author decided to set these requirements to differentiate those participants involved in soybean production from individuals who are not familiar with the soybean crop.

The first step in the snowball sampling tool was to select an initial sampling group. The contact information of this initial sampling group was found through the OSU website. Once the participants were finished answering the questionnaire, the author would ask every participant if they could recommend an individual who met the requirements and would be willing to answer the same questionnaire. If a participant (k) recommended an individual (s), the name of such participant (k) was recorded to be used as a reference. This reference was then used whenever the author contacted an individual (s) without prior notice. This helped to decrease the number of individuals, especially farmers (Pennings, Irwin et al. 1999), denying their participation in this study.

According to Pennings (1999), farmers might refuse to participate or even give unreliable answers to protect their privacy. Using a reference when contacting an participant and assuring their privacy by not publishing the contact's information, might help reduce the probability of an individual declining to participate in a study (Heckathorn 1997). Researchers like Whitley and Ball (2002) suggest that some studies should not focus only in obtaining large sample sizes. According to Whitley and Ball (2002), this might result in a waste of resources, and in some areas like medical school, even unethical to use that many sample. At the end of this process, a total of twenty two participants were sampled by using this sampling tool.

3.2 DETERMINING DISTRIBUTION USING ARENA ® SOFTWARE

Once the sampling population had been selected and the questionnaires had been answered, the next step was to choose a statistical tool able to analyze the results obtained through these questionnaires.

In this study, the author chose the Rockwell Software Arena ® 12 to analyze these results. The input analyzer package from the Arena ® 12 program was used to build a distribution that would represent the results obtained from the questionnaires accordingly. The data was then converted into a text format by using the Notepad ® program. This was done before introducing the results in Arena ® 12.

The first step executed by the program is to create a histogram, allowing the user to visually estimate a distribution. This histogram is built according to the data's frequency. The frequency is based on a lower and upper limit, an estimated average (μ) and a variance (σ). After creating this plot, a normal curve is fitted to this histogram based on the frequency calculated previously. If there are data points located outside the bell curve, there might be a probability that the distribution is non-parametric. A non-parametric distribution is normally used when the researcher does not assume the data being analyzed is normally distributed. Also considered as more robust, the non-parametric distribution is normally used with small sample sizes. Although this distribution is not presumed to be as efficient as the normal distribution, it is also considered to be easier to use and understand (Kaptein, Nass et al. 2010).

3.3 KOLMOGOROV-SMIRNOV (KS) TEST

The K-S is a non-parametric statistical test typically used in statistical software to determine an unknown distribution by using a simulation. This simulation can be used to compare an ideal distribution (Θ) with a hypothesized distribution (ψ) (one sample K-S test). It can also be used to highlight any difference between two hypothesized distributions (two samples K-S test). This tool uses a cumulative distribution function, such that:

$$\Theta = \Pr(X_i \leq x)$$

Where (Pr) represents the probability of a random variable (X_i) falling in a region determined by the variable (x). This tool is also described by the empirical cumulative distribution function, such that:

$$\psi(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

As the value of (n) increases in the previous function, the probability of the $\psi(x)$ will approximate the cumulative distribution function given by(Θ).

Where the summation of the probability of (I) divided by the total number of samples (n) is represented in the following function:

$$I(X_i \leq x) = \begin{cases} 1\ if\ X_i \leq x \\ 0\ otherwise \end{cases}$$

This last equation describes how the summation of the probabilities of (X_i) values falling in a certain range (x) will be equal to 1. If a value falls outside this area, the probability equals 0. These previous equations apply under the assumption that the values being tested follow a continuous distribution. Finally, after performing the K-S statistical test with the Input Analyzer package from Arena 12, this software will produce a KS probability value and a (p) value. This value will be compared with a specific level of significance alpha (this study used $\alpha = 0.05$). If the obtained (p) value is lower than the significance value, the researcher rejects the null hypothesis (H_0). According to Motulsky (2010), the K-S test requires at least 5 or more.

If the researcher is testing one distribution against a known distribution, this means that the distributions are not equal. However, if the (p) value is greater than the significance value, the researcher would accept the null hypothesis (H_o). This would mean that the distribution mean is similar to the known distribution. Usually the known distribution most commonly used is a normal distribution.

3.4 FISHER’S F TEST

The F test will be used to compare the variances of the two samples. Having already explained how this test works, the author decided to describe the hypothesis being tested. The null hypothesis (H_0) will be used to test if the variances of the two samples are equal. The alternative hypothesis (H_1) will be used to test if the variances of the two samples are not equal. Again, the significance value used in this study is ($\alpha = 0.05$) and the degrees of freedom (df) will depend on the sample sizes. Using the values of (α) and the degrees of freedom (df_1 and df_2) the researcher will obtain a specific (p) value. This (p) value will be compared with the value obtained from the F statistical equation to determine if H_0 or H_1 are going to be rejected. The researcher will reject H_0 if the value given by the F statistical equation (f) is greater that the (p) value. However, if the (p) value is greater that the (f) value, the researcher would fail to reject H_0. This means that the sampling variances are equal.

3.5 WELCH’S T TEST

The Welch’s t test is a parametric statistical tool used to compare two samples without assuming equal variances. The Welch’s T test is based on (V) degrees of freedom, an (α) significance test and a (t) value to determine if the two sample means are equal or not. Using the Student T test equation, the researcher will obtain a certain (t) value. This value will be compared with another value given by (α) and (V) from the t table. Welch’s t test is used whenever the sample sizes and variances from the samples are not equal. The following equation describes the Welch’s t test used in this study (Ruxton 2006):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Where N_1 and N_2 represent the different sample sizes, S_1^2 and S_2^2 represent the different variances from two samples being compared, and $\bar{X}_1 - \bar{X}_2$ represent the sample means. The next equation is used to determine the (V) degrees of freedom (Sawilowsky 2002):

$$V = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(\frac{S_1^2}{N_1}\right)^2}{(N_1 - 1)} + \frac{\left(\frac{S_2^2}{N_2}\right)^2}{(N_2 - 1)}}$$

Using a significance value of ($\alpha = 0.05$), the researcher will be able to test the hypothesis. In this study, the null hypothesis (H_0) will be used to test if the means from both samples are equal. This hypothesis will be rejected if the value given by the Student t equation is higher than the (p) value given by (α) and (V). The alternative hypothesis (H_1) will be used to test if the means from both samples are not equal. This hypothesis is rejected if the value given by the Student t equation is lower than the (p) value.

3.6 MANN-WHITNEY U TEST

Once the researcher has determined if the sampling distribution is parametric or non-parametric, the next step is to compare the sampling distribution with the hypothesized distribution. For this study, the author chose Excel ® 2010 to perform Mann-Whitney U test manually. The Mann-Whitney U test is a nonparametric test that compares two independent samples of different size by using a ranking method.

This comparison is made by testing a hypothesis between an experimental sample and a controlled sample. In this study, the experimental sample is being represented by the sampled data and the controlled sample represents the model data from the USDA. The null hypothesis for this study was that there is not a difference between the model data (USDA) and the sampled data. The alternative hypothesis is that there is in fact a difference between the model data and the sampled data. To test these hypotheses, the Mann-Whitney test uses the following equations (Weaver 2002):

$$U_1 = n_E n_C + \frac{n_E(n_E + 1)}{2} - R_E$$

Where U_1 represents the sampled data, n_E is the sample size of the sampled data, n_C is the sample size of the model data, and R_E is the sum of ranks of the sampled data.

$$U_2 = n_E n_C + \frac{n_C(n_C + 1)}{2} - R_C$$

Where U_2 represents the model data, and R_C represents the sum of ranks of the model data. After the researcher has obtained these variables, a U variable is chosen. This U variable will be represented by the smallest value between U_1 and U_2. If the sample sizes (n_E and n_C) are smaller than 5, the U variable obtained previously will be compared with another variable obtained from the U statistic table. If the samples sizes (n_E and n_C) are equal to or bigger than 5, the calculated U value approximates a normal distribution and the following equation is used (Lowry 2011):

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}}$$

The Z value obtained from the previous equation will be compared with a Z value established by the author. In this study, the author is trying to prove with a 95% confidence level that there is a difference between two variables.

This means that the confidence interval according to a two tailed Z test is ±1.96. If the Z value calculated from the Mann-Whitney test falls outside this region, the researcher will have to reject the null hypothesis. However, if the Z value falls between the ±1.96 confidence interval, the researcher will fail to reject the null hypothesis.

3.7 KRUSKAL WALLIS TEST

The Kruskal Wallis is a non-parametric test that compares the variance between three or more distributions. This non-parametric test uses a ranking method to calculate the difference among

the medians being tested. In this study, the author chose Excel ® 2010 to perform a Kruskal Wallis test manually. The inputs were introduced in the spreadsheet, ranked and calculated according to this equation (Green and Salkind 2008):

$$H = \left[\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(N+1)$$

Where H is a variable that represents the distributions being tested, Ri^2 represents the squared value of the summation of ranks in a distribution, n_i represents the sample size of a distribution, and N represents the total amount of observations being tested ($N = n_1,n_2, ..n_k$).. Once this H value has been calculated, it is compared with a (p) value obtained from a Chi-square distribution table.

The value obtained from the Chi distribution table will depend on the degrees of freedom (k-1) and the level of significance α (for this study, the author used $\alpha = 0.05$). If the H variable is greater than the (p) value, the researcher will reject the null hypothesis (H_0). This means that the distributions are different according to the analysis of variance made with the Kruskal Wallis test. If the (H) variable is lower that the (p) value, the researcher will reject the alternative hypothesis.

This means that the distributions are similar and there is not enough statistical evidence to reject the null hypothesis (H_0).

3.8 BIAS

This study was focused on the analysis of soybean yields in the state of Oklahoma, which the author identified as the dependent variable. However, the soybean yield being analyzed could vary according to the independent variables that appear in the following table (Table 2).

Table 2. Independent variables that remained unfixed in this study.

SOURCE OF VARIATION	**INDEPENDENT VARIABLES**
LOCATION	Precipitation or Irrigation Solar Radiation Type of soil Temperature
VARIETY OF SOYBEAN	Seeding rate
AGRICULTURAL TECHNIQUES	Planting date Till or no till technique Planting depth Row width Planting equipment Double crop or single crop

When a researcher targets a more specific population, the number of participants in a study might decrease. To obtain a larger sampled population, the author decided not to consider the independent variables listed previously (Table 2). This decision was made when the author experienced difficulties while trying to contact participants for this study. Keeping this study more generally oriented, allowed the author to target different soybean producers in the state of Oklahoma. Finally, after using a snowball sampling method to contact soybean producers in the state of Oklahoma, the author was able to obtain a sample population of 22 participants

This questionnaire was made U.S. state specific and not geographic location specificand did not account for the independent variables listed in Table 2. However, the author suggests for future studies to analyze results obtained from soybean producers located in specific regions of Oklahoma. After meeting with Dr. Gopal Kakani from the Plant and Soil Sciences at Oklahoma State University on March 7th, 2011, Dr. Kakani mentioned that the precipitation, solar radiation, type of soil and temperature variables are subject to change according to the geographic location of soybean crop. After a short discussion on April 25th, 2010, soybean expert Dr. Chad Godsey from the Plant and Soil Sciences Department at Oklahoma State University agreed with Dr. Gopal Kakani’s in regards to the importance of defining a geographic location when implementing a survey. The following figure (Figure 10) will describe what Dr. Godsey and Dr. Kakani were explaining about the variation of soybean yields in the state of Oklahoma due to different geographic location.

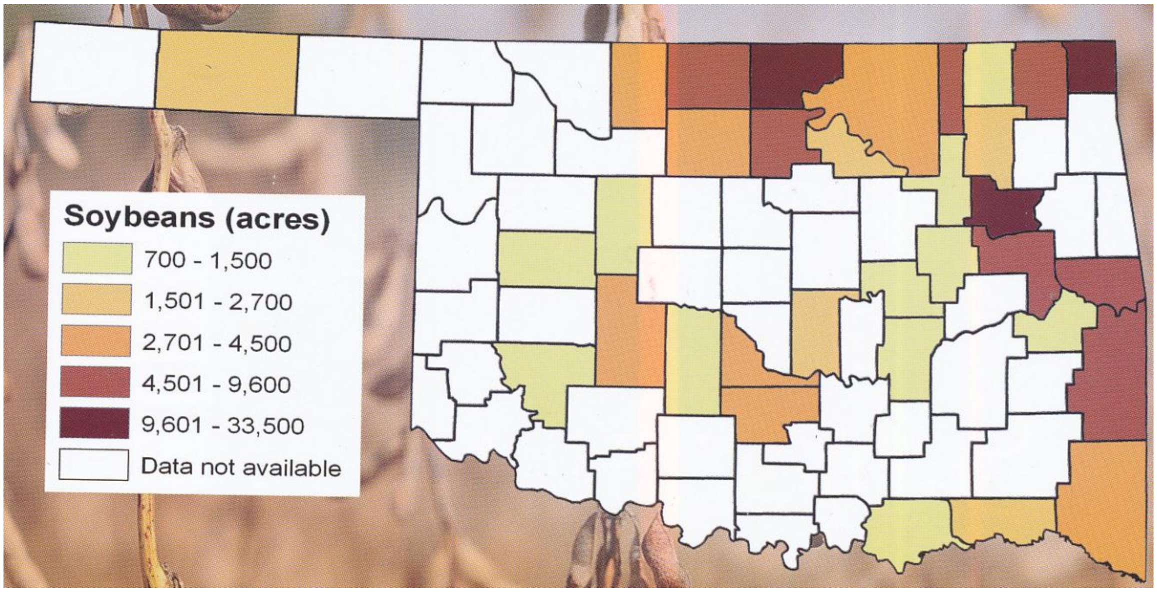


Figure 10. Variation of soybean yields in bushels per acre due to geographic location. Source: Dasnr.okstate.edu

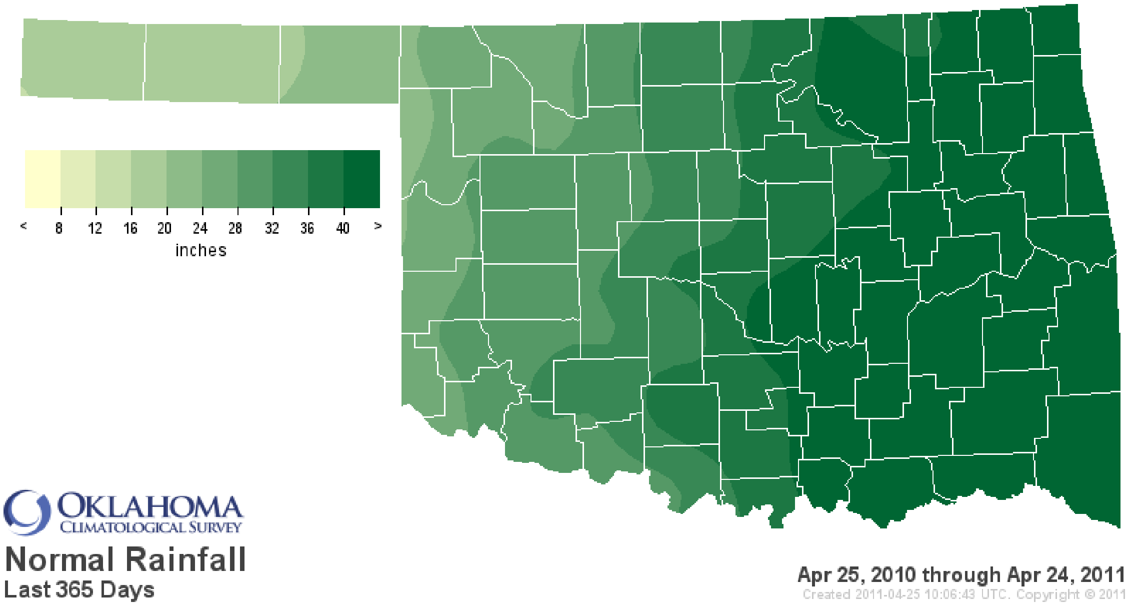
Variation of soybean yields due to different geographic location might be caused by a difference in temperature, water available, day length, variety of soybean, among other factors (Armstrong et al. 2009). However, according to Dr. Chad Godsey, the variation of soybean yields is mainly caused by changes in:

1. Precipitation
2. Temperature
3. Planting date

The next sub-sections will describe each one of these independent variables, unfolding the impact of each variable in soybean yield and the importance of why should they be analyzed in a future study.

3.8.1 VARIATION CAUSED BY PRECIPITATION

“Success or failure depends largely on the weather” (Barnston, Kumar et al. 2008). The increase in size of soybean plants during growth stage depends on the weather (Armstrong et al. 2009), where weather is defined as a description of the atmosphere according to the temperature (hot or cold), moisture (wet or dry), solar radiation (clear or cloudy), pressure, wind, and precipitation (National Climatic Data Center 2009). According to Anderson (2004), soybean yields across the state of Oklahoma are extremely variable due to erratic precipitation. This erratic precipitation is described in the following figure (Figure 11):



<
8
12
16
20
24
28
32
36
40
>
inches
OKLAHOMA CLIMATOLOGICAL SURVEY
Normal Rainfall
Last 365 Days
Apr 25, 2010 through Apr 24, 2011
Created 2011-04-25 10:06:43 UTC. Copyright © 2011

Figure 11. Average precipitation in Oklahoma from April 2010 to April 2011.Source: Mesonet.org

Precipitation is considered an important factor in soybean production because a significant amount of water, equal to 50 percent of the plant’s weight, is required for the soybean plant to germinate. If the amount water is not provided through precipitation, a producer should use an irrigation system to deliver the corresponding amount of water required by the plant. However, if this amount of water is not provided, the producer might expect a decrease in soybean yield.

Research has shown that an approximate 10 percent reduction in water may cause anywhere from 8 to 10% loss in yields (Armstrong, Arnall et al. 2009). The amount of irrigation also varies from 8 inches of water per year in the northeastern part of Oklahoma to 13 inches of water per year in the southwest region of Oklahoma. The following figure (Figure 12) will help illustrate these regions according to the climate zones described by the Mesonet map.

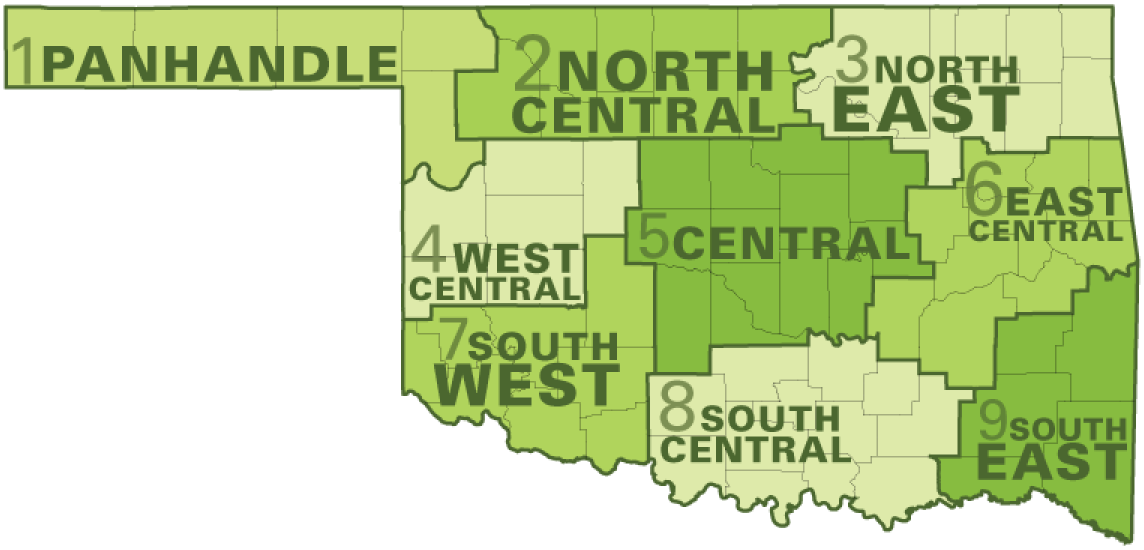


Figure 12. Climate zone map of Oklahoma. Source: Mesonet.org

However, the amount of water provided through irrigation also depends on factors such as the type of soil and the actual temperature (Armstrong, Arnall et al. 2009).The following section will describe the variance in soybean yield caused by temperature.

3.8.2 VARIANCE CAUSED BY TEMPERATURE GRADIENT

Temperature is also considered to be a major influence in the plant’s development (Armstrong, Arnall et al. 2009).

Cold temperatures can slow the seedling emergence and leaf development, while high temperatures tend to enhance the reproductive development. If the temperature drops below 28˚F, the water inside the plant's cells might freeze, causing frost damage. On the other hand, high temperatures (above 95˚F) can reduce the seed quality and the initial germination. For the soybean plant to germinate the temperature needs to reach 55˚F and for the seed to emerge, the temperature should around 60 to 65˚F. These temperatures vary across the state of Oklahoma, and this can be seen in the following figure.

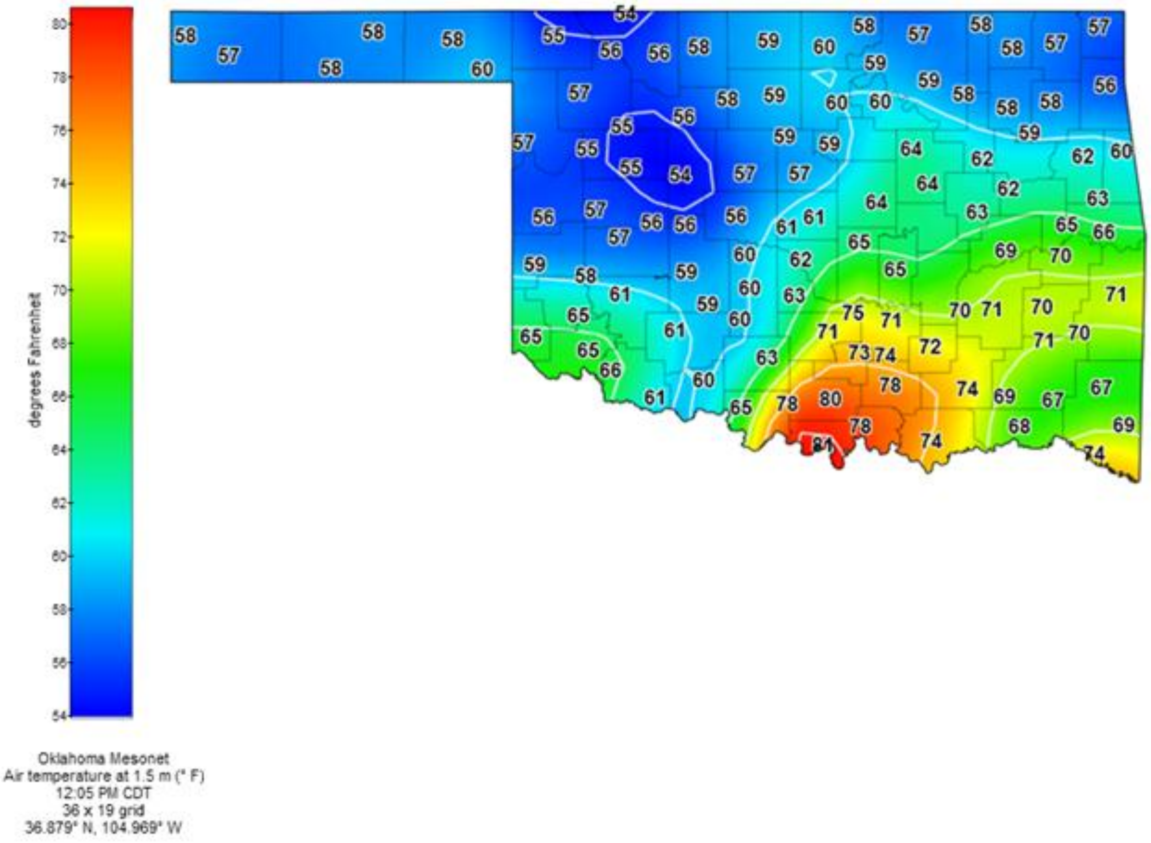


Figure 13. Average air temperature variation in the state of Oklahoma in the month of April 2011

According to Armstrong (2009), the yield loss of soybean is highly dependent on the weather. If the temperature is too hot (above 95˚F), late planted crops could suffer severe losses.

The best recommendation is to plant when the temperature is appropriate for the seed to germinate and emerge, establishing a good stand (Armstrong, Arnall et al. 2009). If the appropiate weather occurs, having the right amount of precipitation and temperature, the other variable that must be taken into consideration is the planting date. This variable will be described in the following section.

3.8.3 VARIATION DUE TO PLANTING DATE

Once the right temperature and the correct amount of water are available, producers might choose to start plating soybean according to the seed variety they have chosen. However, no matter what soybean variety was chosen by a producer, there is still a probability of yield reduction if there are drastic changes in temperature and precipitation. To reduce this probability, Armstrong (2009) recommends having a wide range of planting dates by using different varieties and diversifying to protect the producer against Oklahoma's unpredictable weather. The following table will help describe the range of soybean yields obtained in Goodwell, Oklahoma according to different planting dates (Table 3).

Planting Date	**Yield bu/acre**
5-May	60.5
15-May	72.5
1-Jun	60.1
15-Jun	45.7
2-Jul	33.9
Mean	54.5
Standard deviation	14.94

Table 3. Variation of soybean yields in Godwell, OK caused by planting date. Source: (Kochenower and Scholar 1999)

According to Reddy (2008), an approximate 0.5 bushels of soybean per acre per day are lost if planting date is delayed after the July the 1st. A general recommendation for Oklahoma soybean producers is to start planting in the first two weeks of May and end by June the 15th (Armstrong, Arnall et al. 2009). However, these dates might change according to the weather and the geographic location . To study the impact of the independent variables described previously, simulation models like GIS and ROPGRO allow soybean producers to estimate yields according to the precipitation, temperature and planting date variables specified. In the conclusion of this study, the author will make suggestions of how to analyze the impact of the independet variables described in this section. This will allow future researchers to continue developing this study.

CHAPTER 4

RESULTS

A total of 66 data points were analyzed with the Input Analyzer software from Arena ® software. These 66 data points represent the lower (A), upper (B) and expected variables given by 22 participants in this study. After introducing these 66 data points in the Arena ® 12 software, the Input Analyzer package produced the following frequency plot (Figure 14).

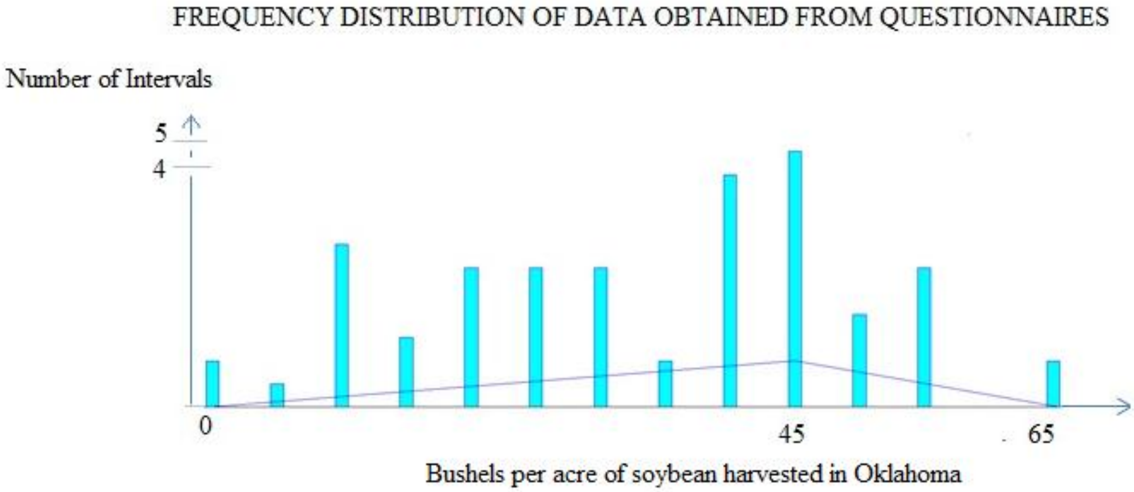


Figure 14. Triangular distribution of data obtained from questionnaires.

After performing a KS test, the triangular distribution was the best fit for the data obtained from the questionnaires. This distribution has a sample mean ($\bar{x}$) of 33.3 bushels per acre with a standard deviation (s) of 16.3 bushels per acre. Table 3 summarizes the frequencies observed in the mode values in this triangular distribution. The percentages described on this table are based on a sample size equal to 22 participants.

Table 4. Frequency distribution of modes in the triangular distribution

Bushels per Acre	**Percentage of Individuals Interviewed**
45	22
40	18
25	14
55	14
30	14
50	10
Other amount	8

Figure 15 describes the normal distribution followed by the values obtained from the USDA Database. These values represent Oklahoma's soybean state average yields from 1961 to 2010:

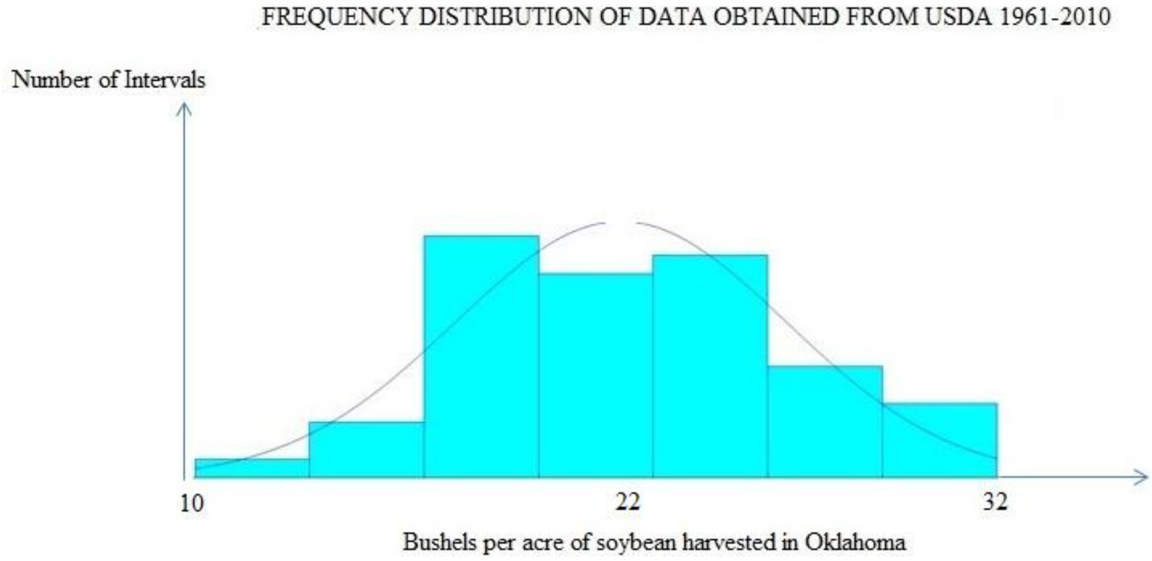


Figure 15. Oklahoma's state average soybean yields from 1961-2010 (USDA 2008).

The data used to build the frequency distribution in Figure 15 appears in the Appendix. This distribution was used as a model to compare it with the triangular distribution and its corresponding subsamples (faculty, farmers and superintendents). An additional graph was built to represent the most recent soybean yields in the state of Oklahoma. This graph (Figure 16) was based on the yields obtained in the state of Oklahoma in the year 2010. The values were obtained from the USDA web site as well:

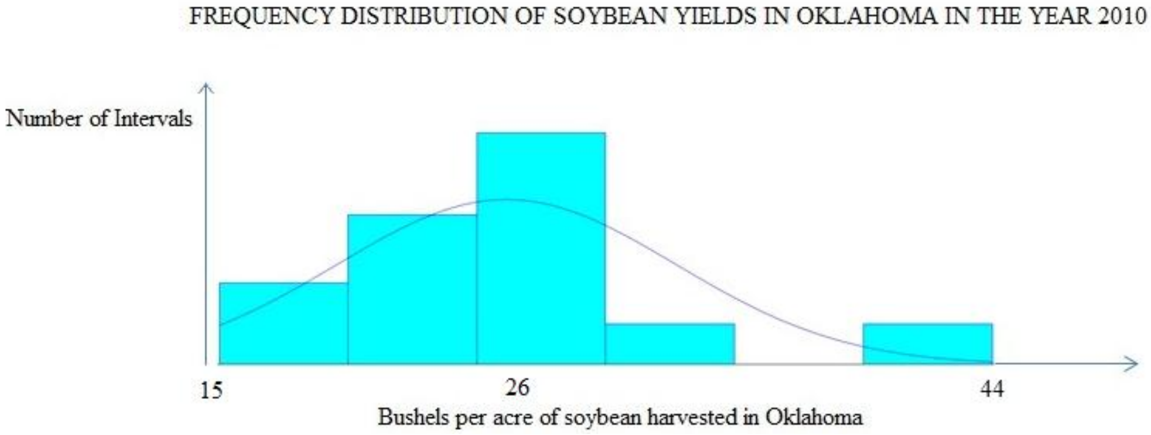


Figure 16. Soybean yields in the state of Oklahoma in the year 2010 (USDA 2008).

The data used to build the frequency distribution in Figure 16 also appears in the Appendix. The three distributions described previously analyze the soybean yield production in the state of Oklahoma taking into consideration all agricultural practices. Agricultural practices refer to irrigated, non-irrigated, single crop and double crop techniques. For this study, the author chose a general setting where no agricultural technique was taken into account to obtain a higher number of participants.

4.1 COMPARISON BETWEEN TRIANGULAR AND NORMAL DISTRIBUTION

The next step was to compare the triangular distribution described by the data obtained from the questionnaires with the normal distribution that describes the soybean yields from 1961 to 2010 according to the USDA web site. This was done by establishing a null and alternative hypothesis:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

According to the Mann-Whitney test, if two samples are taken from a same population in a random way, there should be no difference (Weaver 2002). When using the Mann-Whitney test, the USDA 1961-2010 distribution represents the control group (n_C) and the triangular distribution represents the experimental group (n_E). Also, U_1 will be the variable representing the experimental group and U_2 will be the variable representing the controlled group The results are described in the following (Table 5):

Table 5. Mann Whitney results obtained from comparing the triangular with normal distribution.

	TRIANGULAR DISTRIBUTION (n_E)	**USDA 1961-2010 NORMAL DISTRIBUTION (n_C)**	**SUMMATION**
Size	22	50	72.00
Sums of Ranks	1303.5	1324.5	2628.00
Average Ranks	58.3	26	42.15

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (22)(50) + \frac{22(22+1)}{2} - 1303.5 = 27.5$$

Also, by using the information in Table 4, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (22)(50) + \frac{50(50+1)}{2} - 1324.5 = 978.5$$

After comparing U_1 and U_2, the researcher can observe that U_1 is a smaller than U_2. After choosing U_1 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, instead of consulting the Mann-Whitney test to obtain a probability, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{27.5 - \frac{(22)(50)}{2}}{\sqrt{\frac{(22)(50)(22+50+1)}{12}}} = -6.36$$

If we compare the value -6.36 obtained previously with the -1.96 value given by the Z table when using $\alpha = 0.05$, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

After comparing the triangular distribution obtained from the results of the questionnaires with the normal distribution representing the USDA database, the next step is to find out which of the subsamples (farmers, faculty and superintendent) caused this difference. However, before comparing these subsamples with the model distribution, the researcher must determine what type of distribution these subsamples follow. Even though the researcher has already plotted the data previously and such data was described by a triangular distribution, the author decided to repeat the Arena ® 12 analysis by determining each individual's distribution.

According to Kotz (2004), the triangular distribution could not "reasonably" represent the normal distribution. Also, Johnson (1997) suggested that the comparison between a triangular distribution and a normal distribution would be impossible. However, Johnson (1997) later described how the triangular distribution would possibly represent a beta distribution, without giving sufficient evidence of how a triangular distribution may not be able to represent a normal distribution. Although both these arguments made by Johnson (1997) and Kotz (2004) would

support the result obtained from the previous statistical test, where the null hypothesis stating that there was no difference between the triangular and the normal distribution was rejected, the author decided to test the subsamples (faculty, farmers, superintendents, not so experienced, experienced and experts) that formed the triangular distribution to provide with some statistical results. Finally, if the statistical tools used to test this null hypothesis when analyzing the subsamples is again rejected, then the author would be supporting Johnson's (1997) work with results from these statistical tests.

4.2 COMPARISON BETWEEN FARMER SUBSAMPLE AND NORMAL DISTRIBUTION

The first subsample to be analyzed was the farmer group, and the distribution is the following (Figure 17):

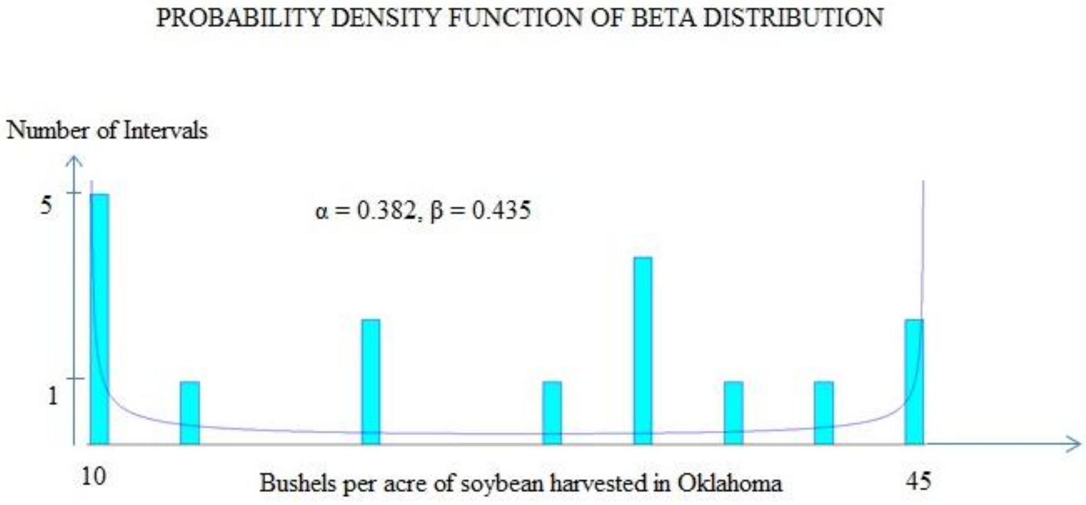


Figure 17. Beta distribution describing the data points obtained from farmers

This previous graph (Figure 17) has a sample mean ($\bar{x}$) equal to 31 bushels per acre with a standard deviation (s) of 17 bushels per acre. According to Sleeper (2007), the value (α) describes the lower part of the previous graph and (β) describes the upper part. The U shape describes a distribution that has a lower limit, represented by the letter A, of 10 bushels per acre, and an upper limit, represented by the letter B, of 45 bushels per acre. This Beta distribution is most commonly used to represent the probability of a variable (x) falling between the intervals (A) and (B), where (α) and (β) are distribution variables that define the curve being plotted. The shape of the Beta distribution will change according to the values of (α) and (β). For this sample, the author performed a Mann-Whitney test to test if there is a difference between the Beta distribution described in Figure 13 and the model data. In this statistical test, (H_0) and (H_1) are trying to test that:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ Beta\ and\ the\ normal\ distribution$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ Beta\ and\ the\ normal\ distribution$$

The results for this statistical test can be observed in Table 6:

Table 6. Mann Whitney results from comparing the Beta distribution describing the farmer subsample with the normal distribution describing the model data.

	FARMERS BETA DISTRIBUTION (n_E)	**USDA 1961-2010 NORMAL DISTRIBUTION (n_C)**	**SUMMATION**
Size	5	50	55
Sums of Ranks	239	1301	1540
Average Ranks	47.8	26.02	36.91

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (5)(50) + \frac{5(5+1)}{2} - 239 = 26$$

Also, by using the information in Table 5, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (5)(50) + \frac{50(50+1)}{2} - 1301 = 224$$

After comparing U_1 and U_2, the researcher can observe that U_1 is a smaller than U_2. After choosing U_1 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{26 - \frac{(5)(50)}{2}}{\sqrt{\frac{(5)(50)(5+50+1)}{12}}} = -2.89$$

If the value -2.89 obtained previously is compared with the -1.96 value given by the Z table when using $\alpha = 0.05$, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

4.3 COMPARISON BETWEEN FACULTY SUBSAMPLE AND NORMAL DISTRIBUTION

Now that the researcher has compared the farmer sampled population, the next subsample to be analyzed will be the faculty members. After running a KS test with the Arena ® 12 software, the distribution that best fitted the farmers subsample was the following (Figure 18):

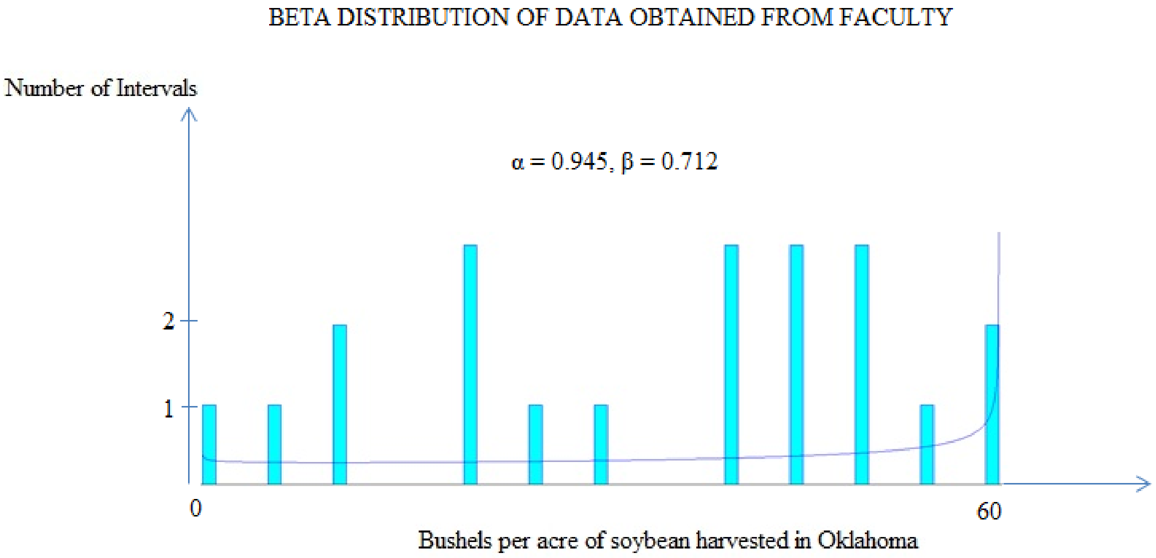


Figure 18. Beta distribution describing the data obtained from OSU faculty members.

Being this statistical test similar to the farmer's test described previously, (H_0) and (H_1) will continue testing for a difference between the distributions by using a Mann-Whitney test, and the results were the following (Table 6):

Table 6. Mann Whitney results from comparing the Beta distribution describing the faculty subsample with the normal distribution describing the model data.

	FACULTY BETA DISTRIBUTION (n_E)	**USDA 1961-2010 NORMAL DISTRIBUTION (n_C)**	**SUMMATION**
Size	7	50	57
Sums of Ranks	372	1281	1653
Average Ranks	53.14	25.34	39.24

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (7)(50) + \frac{7(7+1)}{2} - 372 = 6$$

Also, by using the information in Table 6, the author was able to calculate the value of U$_2$, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (7)(50) + \frac{50(50+1)}{2} - 1281 = 344$$

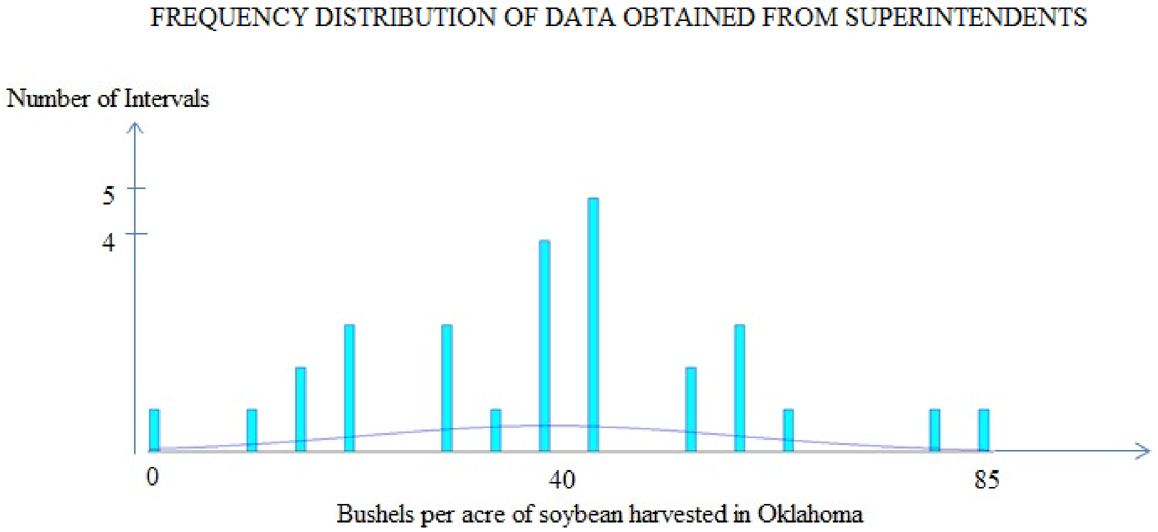
After comparing U$_1$ and U$_{2,}$ the researcher can observe that U$_1$ is a smaller than U$_2$. After choosing U$_1$ as the U value for this test, and also considering that this test has a sample bigger than 5, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{6 - \frac{(7)(50)}{2}}{\sqrt{\frac{(7)(50)(7+50+1)}{12}}} = -4.10$$

If the value -4.10 obtained previously is compared with the -1.96 value given by the Z table when using α = 0.05, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

4.4 COMPARISON BETWEEN SUPERINTENDENT SUBSAMPLE AND NORMAL DISTRIBUTION

Next, the superintendent subsample will be analyzed by following the same procedure. After analyzing the data with Arena ® 12, this statistical software produced the following distribution (Figure 15):



FREQUENCY DISTRIBUTION OF DATA OBTAINED FROM SUPERINTENDENTS

Figure 19. Normal distribution followed by farmer’s data obtained from the interviews

For this specific subsample, the author performed a Welch’s t test since the distribution appears to be normally distributed. However, before applying a Welch’s t test, the author decided to test the variances of the farmer’s subsample and the model data. The F test was used to compare the variances of these two samples. If the variances are statistically different, the author can then choose the Welch’s t test to compare the means of both populations. The results for the F test were the following:

$$F = \frac{S_1^2}{S_2^2} = \frac{169.16}{21.05} = 8.03$$

And the value obtained from the F tables was:

$$F_{(\frac{\propto}{2}, n_1-1, n_2-1)} = F_{(\frac{0.05}{2}, 10-1, 50-1)} = 2.07$$

Being 8.03 a higher value when compared to 2.07, the null hypothesis (H_0) is rejected. This means that there is a statistical difference between the variance of both samples.

Then, the author used the Welch’s t test to compare the mean of the superintendent’s distribution and the mean from the model sample (USDA). The null hypothesis (H_0) and the alternative hypothesis (H_1) are described as:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ means\ of\ both\ samples$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ means\ of\ both\ samples$$

The results of the Welch's t test are described in Table 8:

Table 8. Welch's t test performed on subsample "Superintendents"

	AVERAGE	VARIANCE	SIZE
USDA	21.68	21.05877551	50
SUPERINTENDENTS	40.5	169.1666667	10
T	4.51983		
V	9.45265		

Taking into account the values obtained from the previous Table 8, the result for the (t) value obtained from the following formula would be:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} = \frac{40.5 - 21.68}{\sqrt{\frac{169.16}{10} + \frac{21.05}{50}}} = 4.51$$

Then, the (v) value will result in:

$$V = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(\frac{S_1^2}{N_1}\right)^2}{(N_1 - 1)} + \frac{\left(\frac{S_2^2}{N_2}\right)^2}{(N_2 - 1)}} = \frac{\left(\frac{169.16}{10} + \frac{21.05}{50}\right)^2}{\frac{\left(\frac{169.16}{10}\right)^2}{(10 - 1)} + \frac{\left(\frac{21.05}{50}\right)^2}{(50 - 1)}} = 9.45$$

If the researcher approximates the degrees of freedom (v) to 9, and taking into account a significance level ($\alpha = 0.05$) in a two tailed t test, the value given by the T table (Table 8) equals 2.262. This value is lower when compared to the 4.51 value calculated previously. As a result, H_0 is rejected, and the researcher can conclude according to the Welch's t test, that there is a statistical proof to accept H_1.

Before proceeding with our last set of statistical tests, the author will describe the results obtained so far in the following table (Table 9):

Table 9. Results of statistical tests performed on the sampled population and corresponding subsamples.

Units: (bushels/acre)	USDA 1961-2010	Sampled Population	Farmers	Faculty	Superintendents
Size	50	22	5	7	10
Mean value	21.68	36.36	31	34.28	40.5
Standard Deviation	4.51	10.02	6.51	3.45	13
Distribution	Normal	Triangular	Beta	Beta	Normal
Test	-	MWW*	MWW*	MWW*	Welch's
Result of hypothesis	-	Rejected	Rejected	Rejected	Rejected

* Where MWW stands for Mann-Whitney test

From analyzing the previous table, the researcher can notice the different types of distributions describing each subsample and the sampled population. Also, by using the previous table, the author indicates how the sampled population and its corresponding subsamples have been proven statistically different from the normal distribution that describes the USDA database. That is why the results of the hypotheses appear as rejected, because none of the samples have proven to be statistically equal to the model data.

However, there still remains a last set of experiments that the author decided to realize. These experiments analyze the experience level of the participants disregarding their job description. After classifying the sampled population according to the experience level (Table 1) that was described by the years the participant was related with the soybean crop, the next step was to analyze each subsample individually. Before comparing each subsample with the model data, the author analyzed the corresponding data with the Arena ® 12 software to determine the distribution.

4.5 COMPARISON BETWEEN NOT SO EXPERIENCED SUBSAMPLE AND NORMAL DISTRIBUTION

The first group to be analyzed with the Arena ® 12 software was the “Not so experienced” group, and the distribution was described as (Figure 20):

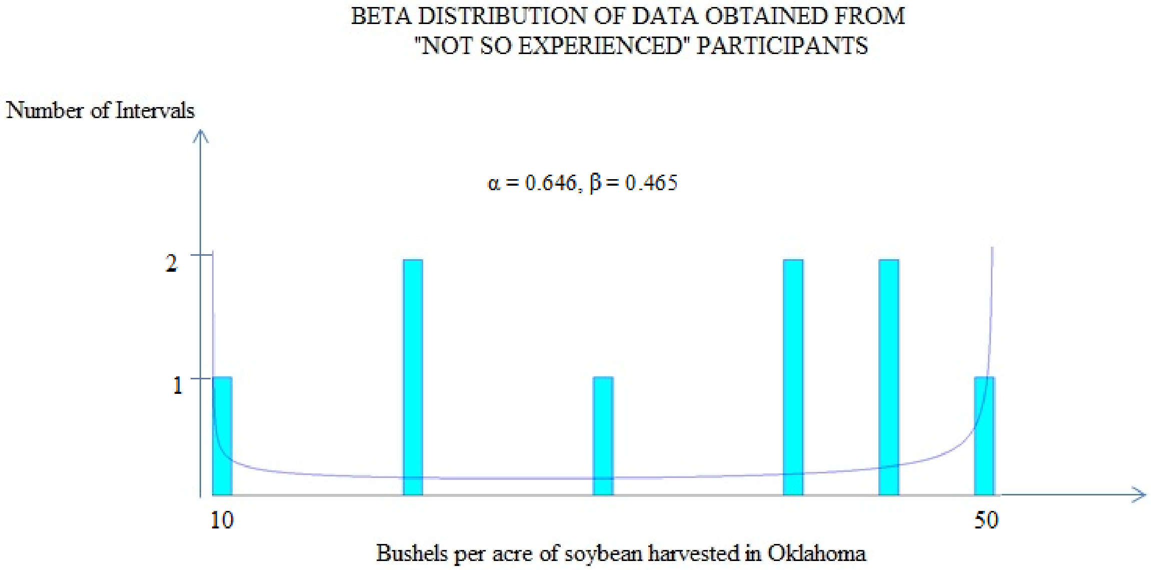


Figure 20. Beta distribution describing the “Not so experienced” subsample

After determining the distribution, the author performed a Mann-Whitney test to test if there is a difference between the Beta distribution described in Figure16 and the model data. In this statistical test, (H_0) and (H_1) are trying to test that:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ Beta\ and\ the\ normal\ distribution$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ Beta\ and\ the\ normal\ distribution$$

The results for this statistical test can be observed in Table 10:

Table 10. Mann Whitney results obtained from comparing the Beta distribution describing the "Not so experienced" subsample with the normal distribution describing the model data

	"Not so experienced" BETA DISTRIBUTION (n_E)	USDA 1961-2010 NORMAL DISTRIBUTION (n_C)	SUMMATION
Size	3	50	53
Sums of Ranks	100	1278	1378
Average Ranks	33.33	25.56	29.44

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (3)(50) + \frac{3(3+1)}{2} - 100 = 56$$

Also, by using the information in Table 9, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (5)(50) + \frac{50(50+1)}{2} - 1378 = 147$$

After comparing U_1 and U_2, the researcher can observe that U_1 is a smaller than U_2. After choosing U_1 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{56 - \frac{(3)(50)}{2}}{\sqrt{\frac{(3)(50)(3+50+1)}{12}}} = -0.73$$

If we compare the value –0.73 obtained previously with the -1.96 value given by the Z table when using α = 0.05, the null hypothesis is not rejected. This means that according to the Mann-Whitney test, there is not statistical difference between these two samples.

To test if this non rejection is statistically possible, and considering the small sample size of the Not so experienced subsample, the author decided to run a second statistical test. Taking into account that the Not so experienced followed a nonparametric distribution, the author decided to prove if this subsample does in fact belong to the same population. To test this hypothesis, the author used the Kruskal Wallis test to test the following hypothesis:

H_0 = There is no statistical difference between the Not so experienced subsample, the model data and the recent data

H_1 = There is a statistical difference between the Not so experienced subsample, the model data and the recent data

For this specific statistical test, the user obtained a different database from the USDA website. The author replaced in this specific statistical test the model data from the years 1961 to 2010 and chose instead the soybean yield estimates from the years 1961 to 2009. This database was chosen to avoid an overlap of data by comparing soybean yields from the same year according to the USDA web page. After analyzing, organizing and ranking the data for the Not so experience subsample, the USDA 1961-2009 data and the USDA 2010 data representing soybean yields, the following table (Table 10) was created (Table 10):

Table 11. Data obtained from the Not so experienced subsample, the model data and the recent data

	"Not so experienced" BETA DISTRIBUTION (n_A)	**USDA 1961-2009 NORMAL DISTRIBUTION (n_B)**	**USDA 2010 NORMAL DISTRIBUTION (n_C)**
Size	3	49	40
Sums of Ranks	259	1795.5	2223.5

Once the calculations have been made, the Kruskal Wallis test will compare the value obtained from a Chi distribution table with the value obtained from the following equation:

$$H = \left[\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(N+1) =$$

$$\frac{12}{92(92+1)}\left(\frac{259^2}{3}+\frac{1795.5^2}{49}+\frac{2223.5^2}{40}\right) - 3(92+1) = 17.98$$

Given a value of α = 0.05, and 2 degrees of freedom obtained from the following equation:

$$Degrees\ of\ freedom(df) = k - 1 = 3 - 1 = 2$$

Where k is equal to the number of samples being analyzed, the Chi square value obtained from the distribution table would be equal to 5.99. If the researcher compares the 5.99 value with the 17.98 value obtained previously, the researcher would reject the null hypothesis because if falls in the rejection region. This means that there is a statistical difference between the three distributions analyzed with the Kruskal Wallis test.

4.6 COMPARISON BETWEEN EXPERIENCED SUBSAMPLE AND NORMAL DISTRIBUTION

Now that the researcher has finished analyzing the Not so experienced sample, the next step is to analyze the Experienced subsample. This Experienced subsample, according to the Arena ® software, is described by the following distribution (Figure 21):

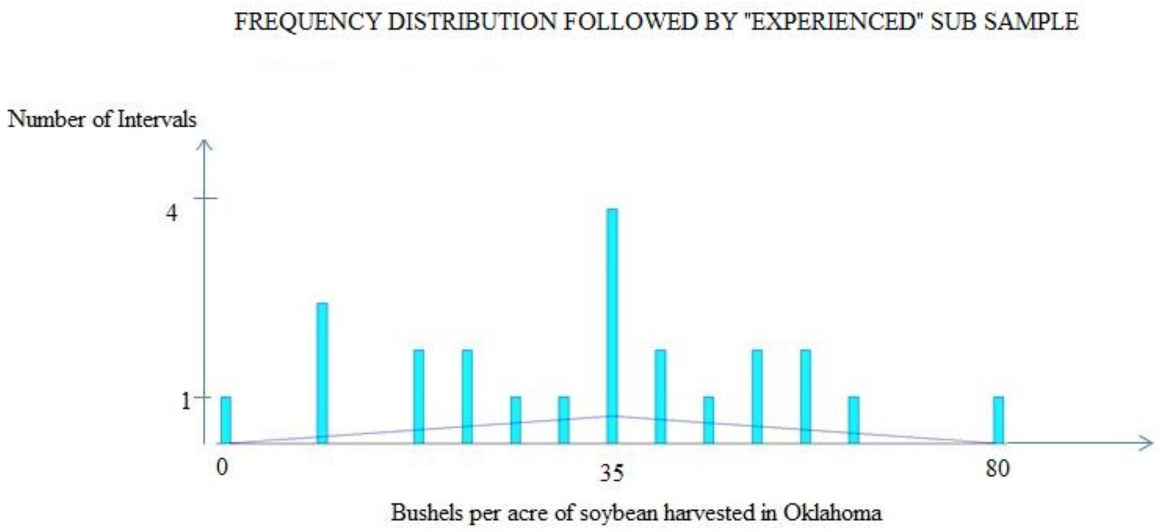


Figure 21. Triangular distribution describing the “Experienced” subsample.

The author performed a Mann-Whitney test to test if there is a difference between the triangular distribution described in Figure 21 and the model data. In this statistical test, (H_0) and (H_1) are trying to test that:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

The results for this statistical test can be observed in Table 12:

Table 12. Mann Whitney results from comparing the triangular distribution describing the “Experienced” subsample with the normal distribution describing the model data.

	“Experienced” TRIANGULAR DISTRIBUTION (n_E)	**USDA 1961-2010 NORMAL DISTRIBUTION (n_C)**	**SUMMATION**
Size	8	50	58
Sums of Ranks	410	1301	1711
Average Ranks	51.52	26.02	38.63

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E + 1)}{2} - R_E = (8)(50) + \frac{8(8 + 1)}{2} - 410 = 26$$

Also, by using the information in Table 11, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C + 1)}{2} - R_C = (8)(50) + \frac{50(50 + 1)}{2} - 1711 = 374$$

After comparing U_1 and $U_{2,}$ the researcher can observe that U_1 is a smaller than U_2. After choosing U_1 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{26 - \frac{(8)(50)}{2}}{\sqrt{\frac{(8)(50)(8 + 50 + 1)}{12}}} = -3.92$$

If the value –3.92 obtained previously is compared with the -1.96 value given by the Z table when using $\alpha = 0.05$, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

4.7 COMPARISON BETWEEN EXPERT SUBSAMPLE AND NORMAL DISTRIBUTION

This statistical test involves the “Expert” subsample, and its distribution is described in the following figure (Figure 18):

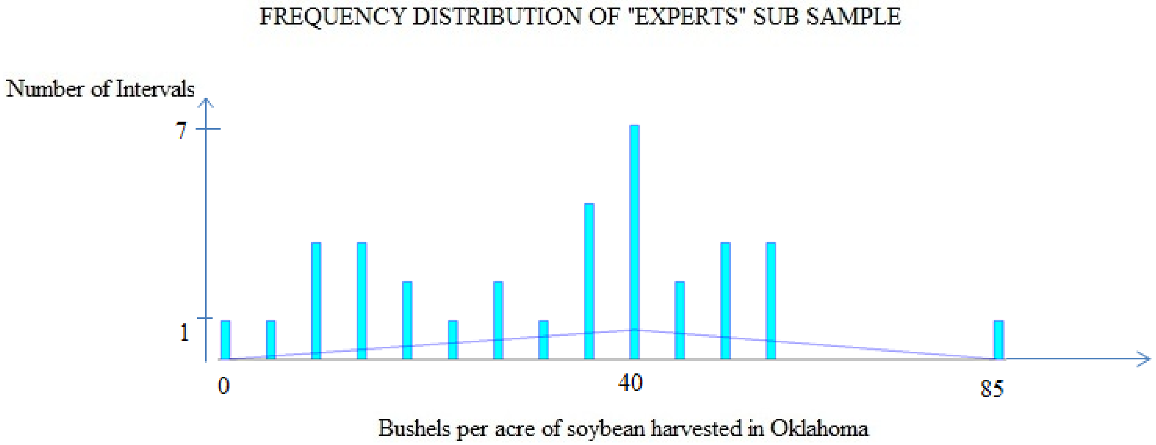


Figure 22. Triangular distribution followed by “Experts” subsample

The author performed a Mann-Whitney test to test if there is a difference between the triangular distribution described in Figure 18 and the model data. In this statistical test, (H_0) and (H_1) are trying to test that:

$$H_0 = There\ is\ no\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

$$H_1 = There\ is\ a\ statistical\ difference\ between\ the\ triangular\ and\ the\ normal\ distribution$$

The results for this statistical test can be observed in Table 13:

Table 13. Mann Whitney results from comparing the Beta distribution describing the “Expert” subsample with the normal distribution describing the model data.

	“Expert” TRIANGULAR DISTRIBUTION (n_E)	**USDA 1961-2010 NORMAL DISTRIBUTION (n_C)**	**SUMMATION**
Size	11	50	58
Sums of Ranks	595.5	1276.5	1872
Average Ranks	54.13	25.91	40.02

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (11)(50) + \frac{11(11+1)}{2} - 595.5 = 20.5$$

Also, by using the information in Table 12, the author was able to calculate the value of U$_2$, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (11)(50) + \frac{50(50+1)}{2} - 1276.5 = 548.5$$

After comparing U$_1$ and U$_{2,}$ the researcher can observe that U$_1$ is a smaller than U$_2$. After choosing U$_1$ as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{20.5 - \frac{(11)(50)}{2}}{\sqrt{\frac{(11)(50)(11+50+1)}{12}}} = -4.77$$

If the value –4.77 obtained previously is compared with the -1.96 value given by the Z table when using α = 0.05, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

Now, the author will summarize the last three statistical tests performed on the not so experienced, experienced and expert subsamples (Table 14):

Table 14. Results obtained from the statistical tests performed on the subsamples pertaining to the experience level

Units: (bushels/acre)	USDA 1961-2009	USDA 2010	Not so experienced	Experienced	Experts
Size	50	40	3	8	11
Mean value	21.61	25.77	33.33	37.5	36.36
Standard Deviation	4.61	6.44	2.88	13.09	9.24
Distribution	Normal	Normal	Beta	Triangular	Triangular
Test	KW*	KW*	MWW** & KW*	MWW*	MWW*
Result of hypothesis	Rejected	Rejected	Non Rejected & Rejected	Rejected	Rejected

* Where KW stands for Kruskal Wallis test

** Where MWW stands for Mann-Whitney test

From the previous table, the researcher can conclude that the subsamples that were tested according to their experience level were found to be statistically different when compared to the normal distribution.

4.8 COMPARISON BETWEEN USDA SOYBEAN YIELDS FROM 1990 TO 2010 AND TRIANGULAR DISTRIBUTION OBTAINED FROM QUESTIONNAIRES

After rejecting the hypotheses in all of the statistical tests described previously, the author decided to compare the triangular distribution obtained from the questionnaires with a different sampled population obtained from the USDA website. This sampled population describes the state average soybean yields in the state of Oklahoma from the year 1990 to 2010. After analyzing the years of experience related with the "Expert" subsample, the author noticed that there was only one individual with 30 years of experience producing soybean. Also, this subsample of "Experienced" participants ranged from 11 to 30 years of experience producing soybean, with an average of 18.5 years and a mode of 20 years. Taking into account the soybean yields obtained in the last 20 years in the state of Oklahoma, the author was able to describe the distribution of the state average soybean yields from 1990 to 2010 according to the USDA database in the following figure (Figure 23):

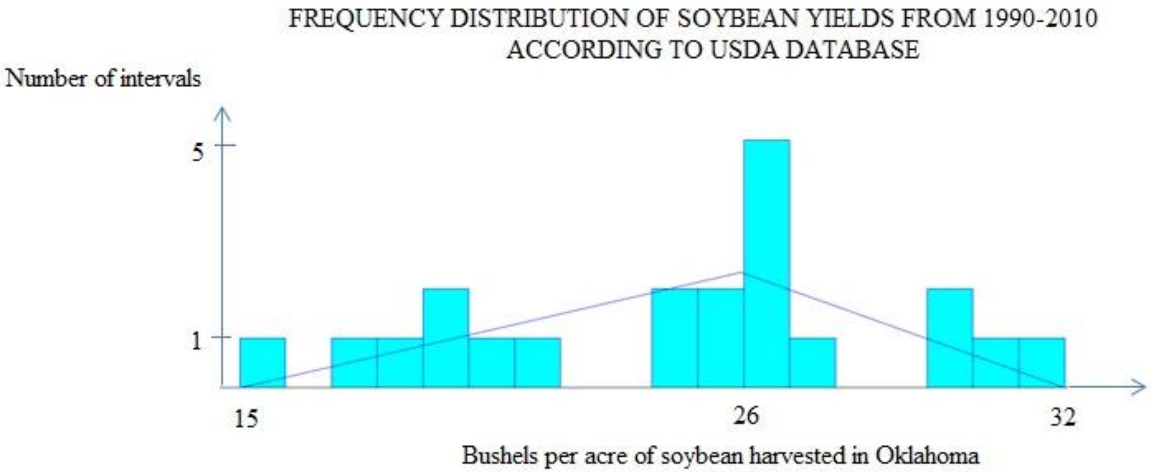


Figure 23. Oklahoma's state average soybean yields from 1990-2010 (USDA 2008).

The data used to build the frequency distribution in Figure 23 appears in the Appendix. The frequency distribution described in Figure 23 followed a triangular distribution. This distribution was then compared with the triangular distribution obtained from the questionnaires. This comparison was done by using a Mann-Whitney test to see if there is a statistical difference between the triangular distribution described in Figure 14 and the triangular distribution described in Figure 23. In this statistical test, (H_0) and (H_1) are trying to test that:

H_0 = There is no statistical difference between the triangular distribution obtained from the questionnaires and the triangular distribution obtained from the USDA 1990-2010 sample.

H_1 = There is a statistical difference between the triangular distribution obtained from the questionnaires and the triangular distribution obtained from the USDA 1990-2010 sample.

The results for this statistical test can be observed in Table 15:

Table 15. Mann Whitney results from comparing the triangular distribution obtained from the questionnaires with the triangular distribution obtained from the USDA 1990-2010 sample.

	TRIANGULAR DISTRIBUTION (n_C)	**USDA 1990-2010 TRIANGULAR DISTRIBUTION (n_E)**	**SUMMATION**
Size	22	21	43
Sums of Ranks	667	279	946
Average Ranks	30.31	13.28	21.80

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (21)(22) + \frac{21(21+1)}{2} - 279 = 414$$

Also, by using the information in Table 14, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (21)(22) + \frac{22(22+1)}{2} - 667 = 48$$

After comparing U_1 and $U_{2,}$ the researcher can observe that U_2 is a smaller than U_1. After choosing U_2 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{48 - \frac{(21)(22)}{2}}{\sqrt{\frac{(21)(22)(21+22+1)}{12}}} = -4.46$$

If the value –4.46 obtained previously is compared with the -1.96 value given by the Z table when using α = 0.05, the null hypothesis is rejected. This means that according to the Mann-Whitney test, there is a statistical difference between these two samples.

4.9 COMPARISON BETWEEN USDA SOYBEAN YIELDS FROM 1990 TO 2010 AND THE EXPERT SUBSAMPLE OBTAINED FROM QUESTIONNAIRES

Finally, the author performed one more statistical test to compare the distribution obtained from the USDA 1990 to 2010 sample with the distribution describing the Expert subsample. The Mann Whitney statistical test was used to test if there is a statistical difference between the USDA 1990 to 2010 distribution and the distribution describing the Expert subsample.

In this statistical test, (H_0) and (H_1) are trying to test that:

H_0 = There is no statistical difference between the triangular distribution obtained from the USDA 1990-2010 sample and the distribution describing the Expert subsample.

H_1 = There is a statistical difference between the triangular distribution obtained from the USDA 1990-2010 sample and the distribution describing the Expert subsample.

The results for this statistical test can be observed in Table 15:

Table 16. Mann Whitney results from comparing the triangular distribution obtained from the USDA 1990-2010 sample and the distribution describing the Expert subsample.

	"Expert" TRIANGULAR DISTRIBUTION (n_E)	**USDA 1990-2010 TRIANGULAR DISTRIBUTION (n_C)**	**SUMMATION**
Size	11	21	32
Sums of Ranks	277	251	528
Average Ranks	25.18	11.95	18.56

Using the previous information, the variable U_1 was:

$$U_1 = n_E n_C + \frac{n_E(n_E+1)}{2} - R_E = (11)(21) + \frac{11(11+1)}{2} - 277 = 75$$

Also, by using the information in Table 15, the author was able to calculate the value of U_2, and the result was:

$$U_2 = n_E n_C + \frac{n_C(n_C+1)}{2} - R_C = (11)(21) + \frac{21(21+1)}{2} - 251 = 211$$

After comparing U_1 and $U_{2,}$ the researcher can observe that U_2 is a smaller than U_1. After choosing U_2 as the U value for this test, and also considering that this test has a sample bigger than 5, the distribution approaches normality (Weaver 2002). Given this previous condition, the researcher would use the following equation :

$$Z = \frac{U - \frac{n_E n_C}{2}}{\sqrt{\frac{n_E n_C(n_E + n_C + 1)}{12}}} = \frac{75 - \frac{(11)(21)}{2}}{\sqrt{\frac{(11)(21)(11+21+1)}{12}}} = -1.60$$

If the value –1.60 obtained previously is compared with the -1.96 value given by the Z table when using $\alpha = 0.05$, the null hypothesis not rejected. This means that according to the Mann-Whitney test, there is not a statistical difference between these two samples.

CHAPTER 5

CONCLUSION

5.1 OVERVIEW OF OBJECTIVES

To conclude this study, the author will refer to the five points described in the beginning of this paper.

1. The author managed to collect data for the triangular distribution with a restricted budget in a considerable amount of time.

The design, outline and length of the questionnaire used in this study may have caused a reduction in the time and money spent when obtaining results from participants. The author decided to use open-ended questions, allowing the participants to provide their own answers without giving any options to choose from. This design may have also reduced the probability of causing an influence in the participant's response. If a researcher provides a list of options to choose from, the probability of obtaining forced results might increase (Richardson Jr. 2002). On the other hand, if a researcher allows the participant to answer a question based on their experience, the results might be considered less biased (Richardson Jr. 2002).

Although some open-ended questions are considered to be more time consuming when compared with closed ended questions (Ahrens and Pigeot 2007), this might have not been the case in this study. If a researcher describes a scenario to reduce possible confusion in a participant before starting with the questionnaire, the researcher might be able to capture a more specific value based on the participant's opinion. In this study, the described scenario was to estimate the state average yield of soybean based on the participant's geographic location and agricultural techniques. Also, the description of a scenario could have also allowed the researcher to reduce the time spent answering a questionnaire.

In a study that involved 100 farmers selected randomly from across the Southern part of the U.S., Pennings (1999) discovered that 35% of the sampled population was not willing to spend more than 5 minutes answering a survey. To encounter this issue, the author designed a questionnaire that might take up to 5 minutes to answer. This could have also raised the response rates of 20% to 30% experienced by Pennings (1999) during his study.

2. After collecting and building a distribution based on 22 samples, the author was not able to prove with a 95% confidence that the triangular distribution represented the soybean yields in the state of Oklahoma.

In a short meeting on April the 7th, 2011, Dr. Carla Goad from the Statistics department at Oklahoma State University suggested to increase the sample size used in this study. Although this may not be the solution for other statistical experiments, Dr. Carla Goad mentioned that the results could be probably influenced by the small sample size. However, if such sample size was increased, there could be a high probability that the amount of time and cost involved in obtaining the necessary data to build a distribution would increase significantly. According to Lehmann (1998), a way of estimating the required sample size in a nonparametric test can be done adding 15% to the number of samples given by a t test, where the equation to calculate the sample size for a two sided t test is (NIST/SEMATECH 2003):

$$N = (t_{\frac{\propto}{2},} t_{\beta})^2 \left(\frac{s}{\delta}\right)^2$$

Where N is the total sample size, $\propto$ is a significance level of $(\propto = 0.05)$ used in this study, β is the probability (β = 0.05) of failing to detect a mean shift from one standard deviation, and δ is used to calculate the shift from the mean value. However, (NIST/SEMATECH. 2003) mentions that this formula requires of certain degrees of freedom, and suggests to use instead the following equation where the standard deviation is assumed to be known (NIST/SEMATECH. 2003):

$$N = (z_{\frac{\propto}{2}}, z_{\beta})^2 \left(\frac{\sigma}{\delta}\right)^2$$

From this equation, the author obtained a sample size of 18, which after adding the 15% suggested by Lehmann (1998), results in 20.7 ∼21 samples. This number does not represent a problem, since the author obtained a sample of 22 participants. Also, Savory (2010) suggests that as the sample size increases, the more the distribution will assimilate the theoretical distribution, in this case, the triangular distribution, and not the triangular distribution.

3. The author was not able to justify the applicability of the triangular distribution in agricultural processes energy and mass flows.

To validate the applicability of the triangular distribution, the author tested a hypothesis where there is no statistical difference between the random variables obtained from a triangular distribution and those obtained from the USDA database when analyzing soybean yields in the state of Oklahoma. After performing a series of statistical analyses with the data obtained from questionnaires answered by participants related with soybean production, the results seem inconsistent with the hypothesis stated previously.

These results show that there is in fact a statistical difference between the triangular distribution obtained from the sampled population and the model data obtained from the USDA web page. The rejection of this hypothesis may have been caused by the level of significance (α = 0.05) used for the studies. However, after having a discussion on April 2, 2010, Dr. Michael W. Smith from the Horticulture and Landscape Department at Oklahoma State University, suggested that this significance level was commonly used in statistical test in the agriculture area. On the other hand, Dr. Smith also commented that the rejection of this hypothesis may have been caused by the small sample size used for this study. However, after analyzing the sample size in the previous point, the author suggests that although the triangular distribution has not proven to be statistically equal from the normal distribution, this statistical tool could still be applied in other areas where the normal distribution is not used. Williams (1992) suggests that the triangular distribution could be used by engineers and managers if it has the capability of representing a 10% value of the best case scenario and a 90% value of the worst case scenario. However, if a researcher is still willing to find a way of comparing the triangular distribution with the normal distribution, the author would suggest furthrer research to be done. One factor that could be taken into account for further research could be a scenario where there is a smaller number of independent variables (Table 2) causing any possible bias in the results obtained from the statistical tests. Also, the researcher might also consider obtaining sample sizes of equal length to use nonparametric statistical tools other than those used in this study.

4. The author was able to create a triangular distribution based on expert's opinion related with soybean yields in the state of Oklahoma.

After interviewing 22 participants related with soybean production in the state of Oklahoma, the author was able to build a triangular distribution using the Arena ® 12 software to plot the results. In order to define the importance or level of experience required to build this type of distribution, the author classified the 22 participants according to three levels of experience. These levels were related with the years of experience the participant had been working with soybean in the state of Oklahoma. After classifying the participants according to their experience level into these three subsamples, each subsample was then compared with the USDA database by testing the hypothesis that there was no a statistical difference between these two samples. The hypothesis was rejected for the three subsamples, meaning that there is in fact a statistical difference between the two samples.

However, the author failed to reject the hypothesis when comparing the soybean yields from the years 1990 to 2010 from the USDA database with the Expert subsample obtained from the questionnaires. Perhaps the experts' opinions were based on more current soybean yields, where different technology and farming techniques are available when comparing to the technology and agricultural techniques used in the 1960's. This is somewhat evident when the researcher notices how the means of the triangular distribution are closer to the more recent USDA yield data (Table 17). As an example, Yang (2009) comments how soybean seed treatment has increased at least 50 percent from the overall soybean planted in the state of Iowa. Hays (2010) suggests that appropriate seed treatment and the variety of seed chosen according to the field's characteristic could maximize the soybean yields. The author recommends further research, where a statistical test compares the triangular distribution with a given distribution that share the same agriculture techniques, geographic location and independent variables that could cause an effect in the soybean yield.

Table 17. Comparison between the Expert subsample, the USDA 1961-2009 sample and the USDA 1990-2010 sample.

	USDA 1961-2009 NORMAL DISTRIBUTION	**USDA 1990-2010 TRIANGULAR DISTRIBUTION**	**"Expert" TRIANGULAR DISTRIBUTION**
Size*	50	21	11
Mean value**	21.61	24.14	36.36
Mode**	18	26	40
Variance**	21	22.82	85.45

* Units in number of participants

**Units in bushels per acre

5. The author was not able to determine if the triangular distribution could be used to describe the variance in energy and mass input streams found in Industrial Ecology projects.

The reason why author was not able to establish the applicability of the triangular distribution to describe the variances analyzed in energy and mass flows in Industrial Ecology projects can be observed in the previous table (Table 16). As the researcher may notice, the variance between the USDA 1961-2009 sample and the USDA 1990-2010 sample is not as different when compared to the Expert subsample. If the variance is going to be used as a variable to determine the boundary limits in a mapping process of an Industrial Ecology project, the author would suggest further research to determine why the variance between the triangular distribution from the Expert subsample and the USDA samples differed so greatly.

An approach to determine the origin of the variance analyzed in Table 16 would be to conduct a survey where the geographic location is taken into account. As described in the Bias section of this study, there is a variety of expected soybean yields in the state of Oklahoma. These expected yields change according to the different geographic locations being analyzed. Each geographic location can be also be differentiated by the amount of precipitation, temperature and plantation dates of the specific region.

Unfortunately, the author was not able to determine the geographic location of the participants due to the objective of keeping the questionnaire anonymous. The author believed that by describing the years of experience, the geographic location and the plantation dates of a soybean producer, this might lead to information identifying the producer. However, if a researcher is able to conduct a study where anonymity is not an issue and contact information of the participant can be described; the author would suggest conducting a questionnaire where the geographic location and the three independent variables (precipitation, temperature and plating date) are taken into account.

Here are four main points that the author suggests could help further researchers in determining the source of variation that was present in this study.

1. Classify participant according to the three main soybean producing areas in Oklahoma: Northeast, North Central and South Oklahoma. This can be done in the snowball sampling process if the researcher asks the person being interviewed to recommend soybean producers from the same area. Once the participants are classified according to their geographic location, the researcher can move on to step 2.
2. Determine if the participant relies on precipitation or if he uses an irrigated system when needed. After analyzing the effects of insufficient water over the expected soybean yields in section 3.9.1, a researcher should be aware of the negative impact caused by the lack of water. As an example, a model made by Reddy (2008) estimated an average loss of 13.39 bushels per acre of soybean yield caused by water stress. Once the participants are classified by irrigated or non-irrigated system, the researcher can continue to step 3.
3. "Temperature is the major factor influencing vegetative development" (Armstrong, Arnall et al. 2009). When the participant is asked for the lowest yield, the researcher might inquire if this low yield was caused by high or low temperatures during that year. If a researcher is able to identify a low yield caused by high or low temperatures, this low yield should also be reflected in other studies analyzing soybean yields. This might explain the cause of a low yield if encountered. Now, the researcher can move on to the next step.
4. Finally, the author suggests asking participants for the planting date. At this point the researcher should already know that after June 15th, the expected yields of soybean start

decreasing. However, a researcher must take into consideration that the planting date is determined by the type of weather. If such weather, mainly described by the temperature and precipitation, is appropriate for a double crop scenario, this will change the planting dates and probably the expected yields. If the weather was not appropriate, the planting date will change, as well as the possibilities for a double crop. Planting dates might allow researchers to identify if a producer used a double crop or a single crop during that year. If such weather did allow the double crop, then the researcher might expect higher yields when compared to other single crop soybean yields.

These are the four main points that could be considered for future studies. After gathering and classifying the information according to these four points described previously, the researcher should be able to test if precipitation, temperature and planting date have an effect on soybean yields according to specific geographic locations in Oklahoma. The three main variables that the author chose for further analysis (precipitation, temperature and planting date) are also taken into consideration in web-based soybean management decision software called *WebGro*. This decision support system allows soybean producers to estimate the effects of different stresses on soybean yields. According to Paz (2004), precipitation, temperature and planting date are the three main environmental constraints that define the expected yields in soybean production. Pedersen (2003) also adds that there are other types of stresses that may affect soybean yields such as pests, herbicide injury, hail. However, it is difficult to estimate the effect that each of these stresses will have on the soybean yield due to the magnitude of compensatory growth and alterations in plant development (Pedersen and Lauer, 2003). If a researcher is interested in analyzing the interactions and effects of different stresses, the author suggests to consult simulation softwares such as CROPGRO and WebGro.

REFERENCES

Administration, I. C. f. M. a. B. (2002). "CPM-Critical Path Method." from http://www.netmba.com/operations/project/cpm/.

Ahrens, W. and I. Pigeot (2007). Handbook of Epidemiology.

Anderson, R. (2004) "Benefits of Crop Sequencing." USDA-ARS.

Armstrong, J., B. Arnall, et al. (2009). "Soybean Production Guide." Oklahoma Cooperative Extension Service Division of Agricultural Sciences and Natural Resources.

Azapagic A. (1999) Life cycle assessment and its application to process selection, design and optimisation. Chemical Engineering Journal 73:1-21. DOI: Doi: 10.1016/s1385-8947(99)00042-x.

Baas, L. (2007). Industrial Ecology as Regional Corporate Sustainability System. International Conference of the Greening of Industry Network Wilfrid Laurier University, Waterloo, Ontario, Canada GIN.

Babonneau, F., O. Klopfenstein, et al. (2010). "Robust capacity expansion solutions for telecommunication networks with uncertain demands."

Bahaman, A. S., L. S. Jeffrey, et al. (2010). "Acceptance, attitude and knowledge towards agriculture economic activity between rural and urban youth: The case of contract farming." Applied Sci., 10: 2310-2315.

Barnston, A. G., A. Kumar, et al. (2008) "Improving Seasonal Prediction Practices Through Attribution of Climate Variability ".

Brown, J. D. (1999). "Standard Error vs Standard error of measurement." JALT Testing and Evaluation **3**(1): 20-25.

Castillo, J. J. (2009) "Snowball Sampling." Experiment Resources.

Castrup, H. (2009). "Error Distribution Variances and Other Statistics." Integrated Sciences Group.

Chen, S.-P. (2007). "Analysis of critical paths in a project network with fuzzy activity times." European Journal of Operational Research 183(1): 442-459.

Cho, J. and H. Garcia-Molina (2003). "Estimating Frequency of Change." ACM Transactions on Internet Technology (TOIT) 3(3).

Cook, J. D. (2009). "Notes on the Negative Binomial Distribution."

Davis, C., I. Nikolic, et al. (2010). "Industrial Ecology 2.0." Journal of Industrial Ecology

14(5).

Dawson, C. (2009). Introduction to research methods: A practical guide for anyone undertaking a research project, How to books.

Department of Environment, F., and Rural Areas, (2010). Agricultural accounts, agricultural prices and farm business statistics. Assessment of compliance with the Code of Practice for Official Statistics, UK Statistics Authority. 72.

Feng, M. G., R. M. Nowierski, et al. (1993). "Binomial sampling plans for the English grain aphid, Sitobion avenae (Homoptera:Aphididae) based on an empirical relationship between mean density and proportion of tillers with different tally thresholds of aphids." Bulletin of Entomological Research 83(2).

Figliola, R. S. and D. E. Beasley (2006). Theory and Design for Mechanical Measurements.

Freund, R. J. and W. J. Wilson (2003). Statistical Methods.

Gao T., Yang H., Liu Y.-l. (2006) Backward tracing simulation of precision forging process for blade based on 3D FEM. Transactions of Nonferrous Metals Society of China 16:s639-s644. DOI: Doi: 10.1016/s1003-6326(06)60269-0.

Garner, A. and G. A. Keoleian (1995) "Industrial Ecology: An Introduction."Pollution Prevention and Industrial Ecology.

Giurco, D., B. Cohen, et al. "Backcasting energy futures using industrial ecology." Technological Forecasting and Social Change In Press, Corrected Proof.

Gomez, K. A. and A. A. Gomez (1984). Statistical Procedures for Agricultural Research.

Goodman, L. A. (1961). "Snowball Sampling." Annals of Mathematical Statistics **32**(1): 148-170.

Greasley, A. (2010). Enabling Simulation Capability in the Organisation, Springer.

Guoping, L. and M. E. Lutman (2006). "Sparseness and speech perception in noise."

Hardaker, J. B., R. B. M. Huirne, et al. (2004). Coping with Risk in Agriculture, CABI Publisher.

Hayashi, K., G. Gaillard, et al. (2007). "Life Cycle Assessment of Agricultural Production Systems: Current Issues and Future Perspectives."

Hays, R. (2010) "Pioneer Offers Tips to Maximize Soybean Success in 2010." Agri Innovations.

Heckathorn, D. D. (1997). "Respondent-Driven Sampling: A new approach to the study of hidden populations." Social Problems **44**(2).

Hye, Q. M. A. (2009). "Agriculture on the Road to Industrialisation and Sustainable Economic growth: An Empirical Investigation for Pakistan." International Journal of Agricultural Economics & Rural Development 2(2).

International Energy Agency. (2000) Hydropower and the Environment: Present Context and Guidelines for Future Action. IEA Hydropower Agreement I.

Ireland, C. (2010). Experimental statistics for agriculture and horticulture.

Irwin, M. E. (2006). "Stratified Sampling."

Israel, G. D. (2009) "Determining Sample Size."

Johnson, D. (1997). "The Triangular Distribution as a Proxy for the Beta Distribution in Risk Analysis." Journal of the Royal Statistical Society. Series D (The Statistician) **46**(3): 387-398.

Jouault, A. (2006). "Determining the Probability of Default of Agricultural Loans in a French Bank."

Joyce, D. (2006) "Common probability distributions."

Kaptein, M., C. Nass, et al. (2010). "Powerful and Consistent Analysis of Likert-Type Rating Scales."

Kapur, A. and T. E. Graedel (2002). "Industrial Ecology."

Keefer, D. L. and S. E. Bodily (1983). "Three-point Approximations for Continuous Random variables." Management Science 29(5): 595-609.

Keefer, D. L. and W. A. Verdini (1993). "Better Estimation of PERT Activity Time Random variable." Management Science 39(9): 1086-1091.

Kim, T.-H. and H. White (2003). "One more robust estimation of skewnedd and kurtosis." Finance Research Letter 1: 56-73.

Kim, Y. J., S. K. Hwang, et al. (2005). "Three-dimensional Monte-Carlo simulation of grain growth using triangular lattice." Materials Science and Engineering: A 408(1-2): 110-120.

Kochenower, R. and R. Scholar (1999) "Soybean Planting Date Test." Soybean Research at OSU, 20.

Korhonen, J. (2004). "Industrial ecology in the strategic sustainable development model: strategic applications of industrial ecolovy." Journal of Cleaner Production 12: 809-823.

Kotz, S. and J. Rene van Dorp (2004). Beyond Beta Other Continuous Families of Distributions with Bounded Support and Applications.

Laboratory, N. R. E. (2011). "Biodiesel Handling and Use Guide, Fourth Edition."

Lane, J. (2009) "Human Ecology article by Pimentel, Cornell colleagues aims to reignite “food vs fuel”, “net energy return” debate on biofuels." Biofuels Digest.

Lehmann, E. L. (1998). Nonparametrics: Statistical Methods Based on Ranks.

Love, G. and J. Goodman (2001). "Overview of Monte Carlo Methodology Used to Estimate Cleanup Project Life-Cycle Cost Uncertainty." Project Performance Corporation.

Lowry, R. (2011). "The Kruskal-Wallis Test for 3 or more Independent Samples."

Lund C., Biswas W. (2008) A Review of the Application of Lifecycle Analysis to Renewable Energy Systems. Bulletin of Science, Technology & Society 28:200-209. DOI: 10.1177/0270467608315920.

Maeda, E. E., P. Pellikka, et al. (2010). "Monte Carlo simulation and remote sensing applied to agricultural survey sampling strategy in Taita Hills, Kenya." African Journal of Agricultural Research 5(13): 1647-1654.

McLaughlin, S. B. and L. A. Kszos (2005). "Development of switchgrass (Panicum virgatum) as a bioenergy feedstock in the United States." Biomass and Bioenergy **28**: 515-535.

Mead, R., R. N. Curnow, et al. (2003). Statistical Methods in Agriculture and Experimental Biology.

Mohammadi, R., A. Abdulahi, et al. (2007). "Nonparametric Methods for Evaluating of Winter Wheat Genotypes in Multi-environment Trials." World Journal of Agricultural Sciences 3(2): 137-242.

Montville, R., Y. Chen, et al. (2002). "Risk assessment of hand washing efficacy using literature and experimental data." International Journal of Food Microbiology 73(2-3): 305-313.

Motulsky, H. (2010). Intuitive Biostatics: A nonmathematical guide to statistical thinking, Oxford University Press.

National Climatic Data Center, N. (2009) "National Climatic Data Center." National Environmental Satellite, Data and Information Service (NESDIS).

Niyaki, S. A. N. and M. S. Allahyari (2010). "Factors Influencing the Adoption of Rice-Fish Farming System in Talesh Region, Iran." World Journal of Fish and Marine Sciences 2(4): 322-326.

NIST/SEMATECH (2003) "e-Handbook of Statistical Methods."

Paulson, E. (1942). "An Approximate Normalization of the Analysis of Variance Distribution."

Paz, J. O., W. D. Batchelor, et al. (2004) "WebGro: A Web-Based Soybean Management Decision Support System."

Pedersen, P. and J. G. Lauer (2003). "Soybean Growth and Development in Various Management Systems and Planting Dates." Crop Science Society of America.

Pennings, J. M. E., S. H. Irwin, et al. (1999). "Surveying Farmers." Review of Agricultural Economics **24**(1): 266-277.

Perlack, R. D., L. L. Wright, et al. (2005). "Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply." Oak Ridge National Laboratory.

Philpott, T. (2006) "An interview with David Pimentel." Grist: A beacon in the smog.

Pimentel, D. and T. Patzek (2008). "Ethanol Production Using Corn, Switchgrass and Wood; Biodiesel Production Using Soybean." Biofuels, Solar and Wind as Renewable Energy Systems.

Qin, J. and R. Lu (2008). "Monte Carlo simulation for quantification of light transport features in apples." Computers and Electronics in Agriculture 68: 44-51.

Rangaswamy, R. (1995). A textbook of agricultural statistics, New Age International.

Ricklefs, R. E. and G. L. Miller (2000). Ecology, W.H. Freeman and Company.

Robertson, T., B. C. English, et al. (1998). Evaluating Natural Resource Use in Agriculture.

Richardson Jr., J. V. (2002) "Open versus Closed Ended Questions In the Reference Environment."

Ruxton, G. D. (2006). "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test." Division of Environmental and Evolutionary Biology.

Särndal, C.-E. and B. S. Jan Wretman (1992). Model Assisted Survey Sampling, Springer.

Savory, P. (2010) "Impact of Sample Size on Approximating the Triangular Distribution." Mathematica Software Demonstration.

Sawilowsky, S. S. (2002). "Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When The Variances Are Different." Journal of Modern Applied Statistical Methods 1(2): 461-472.

Schmitz, A., H. Furtan, et al. (2009). "Agricultural Policy: High Commodity and Input Prices." Agricultural and Resource Economics Review 38(1): 18-35.

Schroth, G. and J. Lehmann (1994). "Contrasting effects of roots and mulch from three agroforestry tree species on yields of alley crooped maize." Agriculture, Ecosystems and Environment 54: 89-101.

Schulz, W. (2007) "The Costs of Biofuels." Chemical and Engineering News **85**, 12-16.

Seager, T. P. and T. L. Theis (2002). "A uniform definition and quantitative basis for industrial ecology." Journal of Cleaner Production 10(3): 225-235.

Seal, H. L. (1949). "Historical development of the use of generating functions in probability theory." 49: 209-228.

Simpson, T. (1757). "An attempt to show the advantage arising by taking the mean of a number of observations in practical astronomy." 64-75.

Sleeper, A. (2006). Desing for Six Sigma Statistics: 59 Tools for Diagnosing and Solving Problems in DFSS Initiatives.

Sleeper, A. D. (2007). Six Sigma Distribution Modeling. Fort Collins, Colorado.

Stein, W. E. and M. F. Keblis (2008). "A new method to simulate the triangular distribution." Elsevier 49: 1143-1147.

Thompson, S. K. (2002). Sampling. New York.

USDA, N. A. S. S.-. (2008). "Quick Stats County Data." from http://www.nass.usda.gov/QuickStats/Screens/faqs.htm#top.

Vose, D. (2000). Risk Analysis - A Quantitative Guide.

Weaver, B. (2002) "Statistics at Square One."

Williams, T. M. (1992). "Practical use of distributions in network analysis." Journal of Operational Research Society **43**(3): 265-270.

Whitley, E. and J. Ball (2002). "Statistics review 4: Sample size Calculations." Critical Care(6): 335-341.

Yang, X. B. (2009) "Soybean Seed Treatment." Integrated Crop Management.

Zou, K. H. and S.-L. T. Normand (2001). "On determination of sample size in hierarchical binomial models." Statistics in Medicine 20: 2163-2182.

APPPENDICES

STATE AVERAGE SOYBEAN YIELDS FROM 1961 TO 2010 (Source: USDA 2008)

Commodity	Year	State	County	Yield	Yield_unit
Soybeans	1961	Oklahoma	State Total	20	bushel
Soybeans	1962	Oklahoma	State Total	17	bushel
Soybeans	1963	Oklahoma	State Total	13.5	bushel
Soybeans	1964	Oklahoma	State Total	16.5	bushel
Soybeans	1965	Oklahoma	State Total	16.5	bushel
Soybeans	1966	Oklahoma	State Total	20	bushel
Soybeans	1967	Oklahoma	State Total	23	bushel
Soybeans	1968	Oklahoma	State Total	21	bushel
Soybeans	1969	Oklahoma	State Total	18	bushel
Soybeans	1970	Oklahoma	State Total	18	bushel
Soybeans	1971	Oklahoma	State Total	21.5	bushel
Soybeans	1972	Oklahoma	State Total	21	bushel
Soybeans	1973	Oklahoma	State Total	22	bushel
Soybeans	1974	Oklahoma	State Total	22	bushel
Soybeans	1975	Oklahoma	State Total	22	bushel
Soybeans	1976	Oklahoma	State Total	22	bushel
Soybeans	1977	Oklahoma	State Total	23	bushel
Soybeans	1978	Oklahoma	State Total	15	bushel
Soybeans	1979	Oklahoma	State Total	23	bushel
Soybeans	1980	Oklahoma	State Total	10	bushel
Soybeans	1981	Oklahoma	State Total	24	bushel
Soybeans	1982	Oklahoma	State Total	18	bushel
Soybeans	1983	Oklahoma	State Total	17	bushel
Soybeans	1984	Oklahoma	State Total	19	bushel
Soybeans	1985	Oklahoma	State Total	23	bushel
Soybeans	1986	Oklahoma	State Total	24	bushel
Soybeans	1987	Oklahoma	State Total	25	bushel
Soybeans	1988	Oklahoma	State Total	18	bushel
Soybeans	1989	Oklahoma	State Total	24	bushel
Soybeans	1990	Oklahoma	State Total	21	bushel

Previous table continues on this page…

Commodity	Year	State	County	Yield	Yield_unit
Soybeans	1991	Oklahoma	State Total	24	bushel
Soybeans	1992	Oklahoma	State Total	27	bushel
Soybeans	1993	Oklahoma	State Total	24	bushel
Soybeans	1994	Oklahoma	State Total	32	bushel
Soybeans	1995	Oklahoma	State Total	20	bushel
Soybeans	1996	Oklahoma	State Total	26	bushel
Soybeans	1997	Oklahoma	State Total	30	bushel
Soybeans	1998	Oklahoma	State Total	18	bushel
Soybeans	1999	Oklahoma	State Total	19	bushel
Soybeans	2000	Oklahoma	State Total	15	bushel
Soybeans	2001	Oklahoma	State Total	19	bushel
Soybeans	2002	Oklahoma	State Total	26	bushel
Soybeans	2003	Oklahoma	State Total	26	bushel
Soybeans	2004	Oklahoma	State Total	30	bushel
Soybeans	2005	Oklahoma	State Total	26	bushel
Soybeans	2006	Oklahoma	State Total	17	bushel
Soybeans	2007	Oklahoma	State Total	26	bushel
Soybeans	2008	Oklahoma	State Total	25	bushel
Soybeans	2009	Oklahoma	State Total	31	bushel
Soybeans	2010	Oklahoma	State Total	25	bushel

SOYBEAN YIELDS FROM 2010 IN OKLAHOMA (Source: USDA 2008)

Commodity	Year	State	County	District	Yield	Yield_unit
Soybeans	2010	Oklahoma	Blaine	20	41.7	bushel
Soybeans	2010	Oklahoma	D20 Combined Counties	20	28.5	bushel
Soybeans	2010	Oklahoma	D20 West Central	20	32.6	bushel
Soybeans	2010	Oklahoma	Alfalfa	40	26.1	bushel
Soybeans	2010	Oklahoma	Garfield	40	17.4	bushel
Soybeans	2010	Oklahoma	Grant	40	23	bushel
Soybeans	2010	Oklahoma	Kay	40	22.1	bushel
Soybeans	2010	Oklahoma	Major	40	43.9	bushel
Soybeans	2010	Oklahoma	Noble	40	18.5	bushel
Soybeans	2010	Oklahoma	D40 Combined Counties	40	26.5	bushel
Soybeans	2010	Oklahoma	D40 North Central	40	22.1	bushel
Soybeans	2010	Oklahoma	Grady	50	24	bushel
Soybeans	2010	Oklahoma	Kingfisher	50	43	bushel
Soybeans	2010	Oklahoma	McClain	50	23	bushel
Soybeans	2010	Oklahoma	Okfuskee	50	21	bushel
Soybeans	2010	Oklahoma	Oklahoma	50	26	bushel
Soybeans	2010	Oklahoma	Payne	50	19	bushel
Soybeans	2010	Oklahoma	Pottawatomie	50	29.9	bushel
Soybeans	2010	Oklahoma	D50 Combined Counties	50	20.3	bushel
Soybeans	2010	Oklahoma	D50 Central	50	25.5	bushel
Soybeans	2010	Oklahoma	Coal	60	15.5	bushel
Soybeans	2010	Oklahoma	D60 Combined Counties	60	27.4	bushel
Soybeans	2010	Oklahoma	D60 South Central	60	27.1	bushel
Soybeans	2010	Oklahoma	Craig	70	25.1	bushel
Soybeans	2010	Oklahoma	Delaware	70	15.4	bushel
Soybeans	2010	Oklahoma	Mayes	70	23.5	bushel
Soybeans	2010	Oklahoma	Osage	70	27.8	bushel
Soybeans	2010	Oklahoma	Ottawa	70	27.7	bushel
Soybeans	2010	Oklahoma	Pawnee	70	22.5	bushel
Soybeans	2010	Oklahoma	Rogers	70	22.3	bushel
Soybeans	2010	Oklahoma	Wagoner	70	26.5	bushel
Soybeans	2010	Oklahoma	Washington	70	27.5	bushel
Soybeans	2010	Oklahoma	D70 Combined Counties	70	28.4	bushel
Soybeans	2010	Oklahoma	D70 Northeast	70	26.3	bushel
Soybeans	2010	Oklahoma	Muskogee	80	31	bushel
Soybeans	2010	Oklahoma	Okmulgee	80	21.3	bushel
Soybeans	2010	Oklahoma	Sequoyah	80	27.8	bushel
Soybeans	2010	Oklahoma	D80 Combined Counties	80	17.3	bushel

Previous table continues on this page…

Commodity	Year	State	County	District	Yield	Yield_unit
Soybeans	2010	Oklahoma	D80 East Central	80	27.4	bushel
Soybeans	2010	Oklahoma	D98 Combined Districts	98	28.9	bushel
Soybeans	2010	Oklahoma	State Total	99	25	bushel

STATE AVERAGE SOYBEAN YIELDS FROM 1990 TO 2010 (Source: USDA 2008)

Commodity	Year	State	County	Yield	Yield_unit
Soybeans	1990	Oklahoma	State Total	21	bushel
Soybeans	1991	Oklahoma	State Total	24	bushel
Soybeans	1992	Oklahoma	State Total	27	bushel
Soybeans	1993	Oklahoma	State Total	24	bushel
Soybeans	1994	Oklahoma	State Total	32	bushel
Soybeans	1995	Oklahoma	State Total	20	bushel
Soybeans	1996	Oklahoma	State Total	26	bushel
Soybeans	1997	Oklahoma	State Total	30	bushel
Soybeans	1998	Oklahoma	State Total	18	bushel
Soybeans	1999	Oklahoma	State Total	19	bushel
Soybeans	2000	Oklahoma	State Total	15	bushel
Soybeans	2001	Oklahoma	State Total	19	bushel
Soybeans	2002	Oklahoma	State Total	26	bushel
Soybeans	2003	Oklahoma	State Total	26	bushel
Soybeans	2004	Oklahoma	State Total	30	bushel
Soybeans	2005	Oklahoma	State Total	26	bushel
Soybeans	2006	Oklahoma	State Total	17	bushel
Soybeans	2007	Oklahoma	State Total	26	bushel
Soybeans	2008	Oklahoma	State Total	25	bushel
Soybeans	2009	Oklahoma	State Total	31	bushel
Soybeans	2010	Oklahoma	State Total	25	bushel