Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Elizabeth Overton Date

Characterizing the underlying statistical distributions of microbial indicators on produce collected from the United States side of the United States – Mexico Border

By

Elizabeth Overton

Master of Science in Public Health

Epidemiology

Juan Leon, PhD, MPH

Committee Chair

Characterizing the underlying statistical distributions of microbial indicators on produce collected from the United States side of the United States – Mexico Border

By

Elizabeth Overton

B.S., University of Oregon – Eugene, 2006

Faculty Thesis Advisor: Juan Leon, PhD, MPH

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
Epidemiology

2013

Abstract

Despite the many health benefits of eating fresh fruits and vegetables, there is a risk of foodborne illness. Fresh produce, since it is consumed raw, never receives a kill step to rid of harmful pathogens. Being able to predict the risk of illness associated with fresh produce is important to prevention. However, before any inferences can be made, the underlying statistical distribution of pathogens, and their associated indicators, needs to be understood in order to make accurate risk predictions. This study assessed the fit of 5 commonly used distributions (normal, lognormal, Poisson, gamma, negative binomial) among 4 indicators (aerobic plate count, coliforms, *Escherichia coli, Enterococci spp.*), sampled from cabbage (n= 109), cantaloupe (n= 42) and cilantro (n= 141), which were collected on the U.S. side of the United States – Mexico border. Distributions were assessed by comparing the Pearson's chi-square values, along with the Akaike's information criterion, to determine which distributions fit each of the indicators. If more than one distribution fit an indicator, the best fitting distribution was determined. Of the 12 different sets of indicator-produce combinations, 10 were found to fit at least one of the assessed distributions. The lognormal fit all 10 of these indicator-produce combinations, while the gamma and negative binomial also fit 6 of the 10 indicator-produce combinations. The normal and Poisson did not fit any of the indicator–produce combinations. For the indicators in which more than one distribution fit, the lognormal was consistently found to have the best fit. This study emphasizes the value in assessing different distributions before making any risk predictions.

Characterizing the underlying statistical distributions of microbial indicators on produce collected from the United States side of the United States – Mexico Border

By

Elizabeth Overton

B.S., University of Oregon – Eugene, 2006

Faculty Thesis Advisor: Juan Leon, PhD, MPH

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
Epidemiology

2013

Acknowledgements:

First, I would like to thank Dr. Juan Leon, whose patience and guidance exceeded any expectations. I could not have picked a better, more encouraging mentor!

Second, a special thank you is also extended to the entire Clean Green team, who collected all of the data. Without their hard work, this study would not have been possible.

TABLE OF CONTENTS

LITERATURE REVIEW

Introduction

Fresh produce consumption has increased dramatically in recent decades [1, 2]. With an increasing emphasis on health, new trends in foods, and increased availability, Americans are more likely than ever before to eat fresh fruits and vegetables [2]. However, there is no kill step to rid fresh produce of the many microorganisms that may reside on them [3], leading to foodborne illness. For example, in 2006, more than 200 people in at least 26 states became sickened after consuming *Escherichia coli* contaminated spinach [4, 5]. In 2008, more than 1,200 individuals fell ill after eating *Salmonella* contaminated jalapeno peppers [6, 7]. With raw produce consumption at an all time high, and an ever more complex food production chain, it is critical to have the capability to detect contamination events before produce consumption.

Sources of Contamination

There are many opportunities for pathogens to make contact with fresh produce both during pre- and post-harvest. During these opportunities, there are a wide array of contamination vehicles, such as animals, water, soil, human handlers, and packing houses, and these will be discussed below.

Much of the food we eat is grown outside, where it is exposed to environmental factors, such as wild animals. Cattle are known to be a primary reservoir for *E. coli* O157:H7 [3, 8], which is a severe foodborne pathogen. In 2006 there was an *E. coli* O157:H7 outbreak that was linked to baby spinach, where the strain was isolated in cattle feces [5]. Following the outbreak, researchers were also able to isolate the strain in feral swine [8]. It is exceptionally difficult to prevent all wildlife from entering growing fields, as fences are not

sufficient to prevent many types of animals [3, 8]. This is especially true when trying to control birds, which have the capability to transfer pathogens from outside sources [9]. Furthermore, it was also noted that once wildlife deposited feces in the field, it might be further spread via lawn mowers and sprinklers [9].

One of the more frequently suspected methods of contamination in the field occurs through surface and irrigation water [3, 9, 10]. The temperatures and pH levels of water sources can be optimal for bacterial survival and growth [10]. Ijabadeniyi *et al.* [10] were able to establish an association between local water turbidity and intestinal *Enterococcus* of surrounding animals. Since water characteristics can influence microbial concentrations, it is essential to test or treat the water before applying it to the fields [10]. Even a small amount of contamination has the opportunity to propagate once it comes in contact with plants [10]. Barak *et al.* [3] found that the surface waters in California, where there is high leafy green production, contained *E. coli* O157:H7 and *Salmonella enterica*. Finally, water is not just used for irrigation; it may also be applied as a pesticide diluent and for cooling [6].

Soil ecology consists of a highly diverse ecological niche that contains countless microorganisms. Normal soil flora includes *Clostridium* species as well as *Bacillus cereus* [9], which can both be pathogenic. Organisms are able to enter soil through a variety of mechanisms, including water, as described above. Another route is through the application of manure fertilizers. The application of manure fertilizers is economical and returns nutrients to the soil [11], but potentially contains harmful human pathogens. One study showed that *Salmonella spp.* persisted in soil for months after it was applied via manure based fertilizer [11]. This emphasizes the point that manure should be treated and applied strategically to avoid excess contamination.

Another point of concern regarding produce contamination is worker hygiene. Farm workers' hands have been shown to contain Norovirus and fecal coliforms [12]. The study population was comprised of pickers, classifiers, and packers. They found Norovirus present on up to 53% of a work group at the beginning of a shift. The subsequently handled green bell peppers, also tested positive for Norovirus. De Roever [13] reviewed fresh produce associated outbreaks prior to 1998. Given the sources of the outbreaks, the conclusion was that hands are a critical risk factor for produce contamination and special attention to handling practices is necessary for Good Agricultural Practices. However, while workers may be aware of good hygiene practices, workers may not have easy access to latrines and hand washing stations [6]. Soon & Baines [14] performed a study focused on training farm workers to wash their hands. They found that to effectively increase hand washing, the training needed to not only be specifically targeted to the workers, but the workers also needed to have easy access to hand washing stations.

The equipment and surfaces in packing houses are yet another source of potential contamination. Of special concern are water baths in which the produce is 'dipped' into. If warm produce is dipped into a cold water bath, tissues contract and draw in water [6]. If the water bath is contaminated, those pathogens have then entered the internal tissues of the produce, essentially protecting the now internal pathogens from surface disinfectants. Surfaces within the packing house are another source of contamination. . Some studies have shown that, overall, packing house surfaces, such as belts, are relatively clean and depend on the type of produce processed on the equipment [15]. Others have indicated that produce that made contact with equipment was more likely to become contaminated [16]. Different results from different studies reinforce the importance of thoroughly considering each step in designing study protocols, as different settings and produce types have different concerns.

Sampling Methods

When designing a method to sample produce for contamination, several key factors may significantly affect the test results. Special considerations include the season sampling is performed, which affects temperature and humidity, and the point between the field and packing sheds where samples are collected. Since slight variations in these variables can yield different outcomes, it is imperative to consider each of these issues when designing a protocol.

Temperature and humidity fluctuate seasonally, which in turn affects microbial growth. Most produce growing periods span several months and seasons. Therefore, it is reasonable to suspect that there might be higher and lower periods of microbial contamination throughout the changing conditions during growing periods. It has been shown that for produce grown near the United States-Mexico border, there were higher levels of contamination in the fall, when compared to winter and spring [16].

The point in production between the farm and packing sheds is likely to influence test results. Samples can be taken from the field, shortly after picking but while still in the field, transport to the packing house, and at different post-harvest stages within packing houses. Each of these steps has unique features that potentially affect the level of contamination. For example, it has been demonstrated, with generic *E. coli* in particular, that microbial concentrations are higher at the final stages of preparation when compared to field samples [16]. In the same study, field samples were also consistently shown to have lower concentrations than boxed samples [16]. These data indicate that packing processes are potentially a significant source of contamination. Steps in processing that are designed to

reduce contamination, i.e. washing, may in fact increase the microbial burden if the water is not properly disinfected [17]. In contrast, other studies have shown that there is no significant change in the quality of certain produce items as they progress through the packing processes [15].

Testing procedures, post sampling:

The goal of performing microbial sampling is to successfully collect all viable microbes and transfer them to a suitable culture medium, in order to provide accurate and representative test results of the microbial levels on produce and surfaces. The first step is to prepare the produce sample for testing. Oftentimes, the preparation process damages the sample as it is homogenized, shaken, or rubbed [18]. Many plants release antimicrobial compounds when they are damaged, therefore while testing for microbial activity the process might actually be killing the very microbes in question [9, 19]. In contrast, if the sample is not damaged as extensively as an attempt to preserve its integrity, there is the chance that a portion of the microbes remain attached to the item. Kim *et al.* [18] compared several methods (pummeling, pulsifying, sonication, & hand shaking). They found that, overall, pummeling and pulsifying were the most effective methods for recovery. Yet, certain items, such as tomatoes, had higher recoveries with methods that were not as damaging (such as rubbing), as they did not release antimicrobial compounds.

Variation *within* samples should also be considered when performing microbial sampling. For example, if there is a wash bin containing produce, it is reasonable to hypothesize that there are different degrees of contamination when comparing the top of the bin to the bottom of the bin. This potential scenario was tested with bagged baby spinach

and romaine lettuce. Researchers compared spinach and romaine lettuce samples taken from the top of the bag to samples taken from the bottom of the bag [20]. The study showed that in some cases there were higher counts at the bottom of the bags, and rarely were the counts higher in the top of the bag. While the data was not consistently significant, the authors concluded that they could not reject the hypothesis that there is sample variation *within* the bags.

The plate count is subject to the technician's technique and interpretation [21]. Different technicians might arrive to unequal final counts, as the counts are only as accurate as the skill of the technician. To assist with variability, plate counts are often done in duplicate or triplicate and sometimes by different operators. By replicating plate counts, the effect is an overall reduction in sampling variance and total variances [22]. It is often unknown to what magnitude the initial microbial population is, so it is useful to quantify several serial dilutions. In turn, more colonies and plates are counted, and a weighted average can be obtained. Serial dilutions have the added benefit that the technician is better able to 'catch' the readable plates, but there can be increased error due to increased diluting and sampling [22].

Standard acceptable plate counts are between 30 and 300 colony forming units (CFUs) [21]. The terminology 'colony forming unit,' or CFU, takes into account that visible colonies of bacteria (or yeast and fungus) might be formed by more than one cell, or that smaller colonies have grown into each other and appear as one [21]. Counts greater than 300 CFUs lead to overcrowding and a decreased ability to distinguish between colonies, and might inhibit further growth. Counts less than 30 are potentially inaccurate and not representing higher counts [21]. In the event that the plate counts yield results outside the

parameters, it is suggested the results should be noted as greater or less than the limits of detection [21]. Zero counts are discussed later, but do not necessarily represent a complete absence of the microorganism, but possibly a concentration below the limit of detection, and are often replaced with imputed values. It is infrequently described in the methods sections of peer-reviewed articles as to how these values were treated and why.

Indicator Organisms

Indicator organisms are often used in place of an actual pathogen. Ideally, when testing produce for human pathogens, the pathogen itself would be directly enumerated. However, the pathogens in question are often rare, difficult to culture, and not evenly distributed [23]. This is an issue because some pathogens are infectious even at low doses and need to be detected in some way. As an alternative, organisms that theoretically co-exist with the actual pathogen serve as surrogates, and are referred to as indicator organisms. Despite the value of indicator organisms, there are also drawbacks that limit the usefulness of the data.

The presence of an indicator organism does not identify the source of contamination [23]. Indicators may come from a variety of hosts, whereas the interest and concern might be human or animal contamination. It may also be unclear whether or not the indicator is able to replicate in the environment in the same manner as the actual pathogen. Instead, the resulting enumeration may indicate that the indicator grew in the sample, rather than growth from the environment or an animal source [23]. Some studies have shown a negative correlation between fecal indicators and viral contamination [12, 24]. Indicators need to be carefully selected. Some research has shown that if the true pathogen is anaerobic, but

detected with an aerobic indicator, an aerobic test will represent a lower concentration than actually exists [10]. Therefore, aerobic counts are not necessarily indicators of pathogens that may be anaerobic. Busta *et al.* [25] suggest that an indicator is not appropriate if it exists when the pathogen is absent, or does not exist when the pathogen is present, and is only appropriate if it grows similarly to the pathogen.

Common indicators for human pathogens include total aerobic plate counts, *E. coli*, *Enterococci spp*, and coliforms. The total aerobic plate does not differentiate between different types of aerobic bacteria; it provides a count for a broad spectrum of pathogenic and non-pathogenic microorganisms [26]. Since the plate count is not providing actual counts of pathogens, the plate count cannot be used as a measure of food safety [26]. Instead, the aerobic plate count provides a more general idea of the overall quality. *E. coli, Enterococci spp,* and coliforms are detected as a means to identify fecal contamination. These organisms have long been used as indicators of fecal contamination. *E. coli* was first used as a fecal indicator in water in 1892, and by the early 1900s *Enterococci spp.* and coliforms were being used as fecal indicators in various food products [27]. However, there are some concerns with relying on these indicators as representatives for fecal contamination. Kornackie *et al.* [27] bring up the issues that these organisms can live outside of warm-blooded intestinal tracts, can live in the environment, can become normal flora in food processing settings, and can grow in food products. Furthermore, establishing acceptable limits of fecal coliforms can be difficult. A lack of correlation is seen between many fecal indicators (*Enterococcus spp.* and *E. coli*) and actual levels of fecal contamination in food products [28]. This is also true in water testing, where fecal indicators and actual pathogens in water have shown to have weak correlations [29]. Instead of assuming that the indicator is either useless or highly accurate, it has been suggested to consider the presence of indicators as a *sign* of risk [25].

Some microbes are able to be linked to a specific host through the use of microbial source tracking (MST). MST is based on the concept that certain markers are specific to certain hosts [23]. There are multiple methods to accomplish MST, such as culture-based, chemical, and molecular techniques. Not all microbes can be tested with MST, so it is not inclusive for all pathogens. However, it is an additional tool that may prove to be successful in certain studies.

Contamination Distributions

The physical distribution of microbes between and within food items will have an impact on the test results [4, 30, 31]. Additionally, the statistical distribution that the data is assumed to fit will in turn influence risk inferences and predictions [32]. Understanding how contaminants are distributed physically, and how to statistically represent these distributions, has far reaching implications for food safety predictions.

Physical distributions

Different locations within a production chain, and produce item, are likely to have different distributions of microbes [4, 31, 32]. Therefore, samples collected from different locations cannot be expected to have the same microbial concentrations. Likewise, depending on the portion of the produce item sampled, the concentration of microbes may vary *within* the produce item selected. It is possible that microbes are uniformly distributed throughout the item or surface, but it is more likely that concentrations are clustered [33]. Many produce types have natural defenses against pathogenic microbes, such as natural antimicrobials and surface morphology characteristics [19]. As these characteristics are not uniform throughout the entire produce type, the distributions of microbes will also vary.

Some of the factors that contribute to growth and death of microbial populations are known and measured, but there are also many factors that remain unknown [4].

Food and water microbial count data frequently display irregular and random fluctuations [4], as there are many variables that lead to the heterogeneity seen in microbial food items [31]. This should not be surprising since there are many factors involved in microbial growth. It seems exceptionally unlikely that one would observe a random distribution of microbes within a sample [34]. Jongenburger *et al.* [33] suggest using the ratio of variance to mean to indicate the degree of clustering (a high ratio indicates high levels of clustering) as a means of estimating the data's heterogeneity.

It is important to remember that microbial testing is not completely accurate, and might not detect any indicators when they are in fact present. Closely related to the accuracy of results is the sample size. Gonzales-Barron & Butler [34] clearly explain how sample size not only affects the observed mean and variance from the sample, but also the observed prevalence of the microbes tested. Therefore, test results need to be treated as a representation of the true data. Despite even large sample sizes, it is common to not detect any microbes when testing for rare indicators such as *E. coli*. The most frequent result may be zero plate counts. The zero count can either represent an absence of the pathogen, or that the pathogen existed in a concentration below the limit of detection [33]. For example, perhaps one took a 10 gram sample from a 200 gram piece of produce. On this produce item, there might only be 18 CFU of a particular pathogen (which can be an infectious dose for some pathogens), meaning an average concentration of approximately less than 1 CFU per 10 gram sample. Therefore, there is a chance that by only testing 10 grams one fails to detect any of the indicators, but that does not mean that there was an absence of the

indicator, rather it existed below the limit of detection. So, even though one had a plate count of zero, one cannot report with confidence that there was a complete absence. Zero plate counts are a significant issue and worthy of a separate review, but needs to be considered when analyzing data [21, 35, 36]. Sometimes these counts are handled by imputing data, or leaving them as a zero value.

Statistical Distributions

Historically, the lognormal distribution has been used to describe microbial data from food [31, 32]. However, the lognormal distribution is most suitable for uniformly distributed high concentration values, with no zero counts [32, 33]. Many indicators and pathogens found in food, especially fresh produce, may not exist in high concentrations. Instead, they occur with many sporadic numbers between zero counts [31, 33], meaning that the lognormal distribution may be inappropriate for such data.

There are several issues to consider when fitting a distribution to model plate counts. First, the events in the right tail of the distribution are the most significant in terms of public health impact [4, 33]. Second, the number of microbes present in a sample is a discrete value. It is impossible to have a fraction of a microbe, as is often represented in concentration data. However, concentration data can be represented with a continuous distribution. Third, microbial populations can grow to large sizes, but can never be less than zero [4].

In addition to the lognormal distribution, other distributions that have been investigated for food sampling data include the normal, Poisson, gamma, negative binomial, and Poisson-lognormal distributions [33]. The normal distribution is continuous, does not allow for zeros, and is symmetric. This is not ideal for food microbiology, as a distribution should allow for zero counts, accommodate discrete counts, and is rarely symmetric. Poisson

distributions allow for zero counts, model discrete values, and subtypes are able to accommodate for over-dispersion. The gamma distribution, defined by the parameters scale and shape [32], is another continuous model that does not allow for zero counts, and is not suitable for low counts but can be used as a generalizing distribution for the Poisson distribution [33]. Discriminating between the lognormal and gamma distribution is oftentimes difficult as they are very similar ([32], reviewed in [37, 38]). The negative-binomial distribution arises when the gamma distribution generalizes the mean of a Poisson distribution, and is especially useful when the distribution is over-dispersed, as the negative binomial has additional parameters to accommodate this situation [33]. The negative-binomial allows for zero counts and discrete values, while approximating the lognormal. The Poisson-lognormal arises from when the lognormal distribution generalizes the mean of a Poisson distribution [33]. This type of distribution is valuable for food safety data since it allows for zero counts, discrete values, approximates the lognormal, and may be appropriate for a mixture of distributions [33]. Gonzales-Barron & Butler [31] found that the Poisson-Lognormal distribution was the best representation for low microbial counts with zero values.

To actually determine the best distribution, Jongenburger *et al.* [33] state that fitting the actual observations is necessary, not just considering theoretical assumptions. They provide these guidelines when considering which distribution to fit:

1) The Normal, Lognormal, and Gamma distributions typically model continuous data, such as concentrations.
2) The Poisson, Zero-Inflated Poisson, Negative Binomial, and the Poisson-Lognormal are used for discrete distributions.

Given the above considerations, Jongenburger *et al.* [33] suggest 5 criteria to assess the underlying distribution of microorganisms:

1) The distribution should be non-negative.
2) The distribution should allow for zero values.
3) The distribution should be discrete.
4) The distribution should reduce to the Poisson distribution.
5) The distribution should approximate the lognormal distribution at high numbers.

Using these criteria, researchers should see which distributions best fit their data. Once the proper distribution is fit, better inferences can be made.

While there has been some research, as described above, on characterizing the underlying microbial distributions in food microbiology, there has been no research within the area of fresh produce contamination. As described, fresh produce is an important part of the American diet and has a history of causing illness. Characterizing the contamination distributions is the first step to making risk predictions, and further statistical inferences.

Needs

There is a need to characterize the underlying microbial distributions in fresh produce, where zero counts and heterogeneous data can be expected. In addition to there being relatively little literature available on the underlying distributions of microbes on produce, there is also a lack of microbial testing of produce at the farm level. However, in order to statistically model the distributions, these characterizations need to be performed [33, 34]. By first understanding how to interpret test results, further decisions and protocols

can be developed that may eventually lead to more frequent testing as a means to assess associated risk due to consumption.

Goals and Aims

Goals:

- To assess the fit of 5 different statistical distributions (normal, lognormal, Poisson, gamma, negative binomial) for the concentrations of 4 indicators (aerobic plate count, coliforms, *E. coli*, *Enterococci spp.*) on cabbage, cantaloupe, and cilantro samples collected from the U.S. side of the United States – Mexico border.

Aims

- To determine which, if any, of the distributions fit the data, decided by the Pearson's chi-square statistic.

- To determine which, if any, of the 5 assessed distributions has the best fit for each of the indicator – produce combinations.

Significance

If the underlying distributions and contamination patterns of microbial contamination occurring at the field and farm level were better understood, there would be more incentive for a farmer to test produce before it leaves the packing shed. From the perspective of farmer growing in the United States-Mexico growing region, microbial testing is rather expensive. If it is not even established as to how to treat the resulting data, sampling

procedures are less efficient and the test results have little meaning. At this point, there is little motivation for a farmer to initiate testing. However, knowing how to model the data with appropriate distributions means that better risk predictions can be made, and in turn provide more incentive for a farmer to test their produce for contamination.

Role in thesis

Data was collected by the Clean Greens team (Universidad Autonoma Nuevo Leon, North Carolina State University, Emory University). Further details can be found in [15]. Data analysis was performed by Elizabeth Overton, with assistance from Dr. Robert Lyles and Dr. Peter Teunis (Emory University). The writing of the thesis was performed by Elizbeth Overton, with assistance from Dr. Juan Leon (Emory University).

INTRODUCTION

Foodborne illness outbreaks associated with fresh produce consumption have increased in recent years (reviewed in [39]). Despite the many nutritional benefits of fresh fruits and vegetables, there is also an inherent risk of illness in consuming foods that are not treated (either thermally or chemically) to rid of harmful pathogens [3]. Fresh fruits and vegetables are especially prone to contamination since there are many opportunities for pathogens to come in contact with produce through soil, water, animals, and workers (reviewed in [16, 39]). The result has been large outbreaks that generate serious illness, and even death. Coupled with the effects on consumers is the detrimental effect an outbreak has on a farmer (reviewed in [40]). Being able to predict and detect contamination of fresh produce prior to consumption is paramount to protecting the health of consumers and farmers worldwide.

Obtaining representative samples that contain enough information to make valid risk predictions is particularly difficult, as the pathogens are often relatively rare, and not necessarily homogenously distributed. Many of the pathogens that cause serious illness (i.e. *Escherichia coli* O157:H7) exist at very low prevalence levels [41, 42], and are challenging to detect [23]. Therefore, microbial indicators with similar physical properties are used, and even these organisms are often difficult to detect [23]. Consequently, testing procedures often indicate an absence of these indicators on the tested sample, even when they are present in the larger sample [33]. Furthermore, the spatial distribution of indicators is not necessarily homogeneous, instead they tend to be clustered and distributed irregularly within the item [30]. This is likely to due to the fact that there are many variables affecting the distribution, including environmental fluctuations [16] and physical characteristics that alter

the ability for pathogens to persist [18]. For these reasons, further investigation of how microorganisms are physically and statistically distributed in the larger sample needs to be examined, prior to performing any modeling or risk predictions.

Before any models or predictions can be made, it is essential to identify the underlying statistical distribution of an indicator. Several studies have examined the underlying microbial distributions in meat [34], dairy [4], and drinking water [43]. They have found that the lognormal distribution, which is traditionally used for microbial data, is not always appropriate as it fails to properly model data with an excess of zero counts. These zero counts might actually be values below the limit of detection, and inaccurately indicating a complete absence of an indicator. Instead, it has been recognized that distribution fitting procedures should be performed to determine a distribution that is capable of accommodating the features of data involving rare organisms. These features are not limited to, but include, an excess of values below the limit of detection, sparse data, and heavily skewed results [32, 34]. Possible distributions that can be assessed are the normal, lognormal, Poisson, gamma, and negative binomial. These 5 distributions are common for modeling food safety data (reviewed in [32]). As a consequence of sparse counts, it has also been suggested that the sample size needs to be much larger than expected, maybe as high as 100 samples in the case of drinking water [43]. Despite the body of research on this topic for other food types, fitting the underlying statistical distributions has not been performed with fresh produce. Assessing the underlying statistical distributions is a necessary step in order to make accurate risk predictions, which can thereby decrease the risk of illness and adverse effects on farmers.

Thus, to address these needs, the goal of this study is to assess the fit of 5 statistical distributions (normal, lognormal, Poisson, gamma, negative binomial) in fresh produce samples collected between November 2002 and November 2004, from the U.S. side of the United States-Mexico border. The produce samples were collected as part of the Clean Greens I-II study, performed by Emory University and North Carolina State University. A subset of the samples, cabbage, cantaloupe, and cilantro, were fit to the five different distribution types using proc genmod in SAS 9.3 (SAS Institute Inc., Cary, N.C.). The results indicate that the lognormal distribution has the best fit of the 5 assessed distributions. The findings of this analysis can be used to make further risk predictions and distribution assessments for this data set. It may also highlight the importance of distribution fitting procedures for future studies on fresh produce contamination.

METHODS

Sample Collection

Produce samples (n=490) were collected from 8 packing sheds on the U.S. side of the United States – Mexico border, between November 2002 and November 2004. The analysis included 13 different produce types (broccoli, cabbage, cantaloupe, celery, Swiss chard, cilantro, collards, curly parsley, dill, kale, parsley, root parsley, turnip greens). Within the packing shed, produce was sampled from 8 locations (bin, box, conveyor belt, dump tank, merry-go-round, rinse cycle, wash tank). Samples were aseptically collected in duplicate, at 150g per sample.

Sample Processing

Following sample collection, samples were packed on ice and shipped overnight to the Department of Food Sciences at North Carolina State University. All samples were processed within 24 hours of collection. Samples were tested for total aerobic bacteria (APC), total coliforms, total *Enterococcus*, and total *Escherichia coli.* Samples were divided into 25g sub-samples, and diluted 1:10 in 0.1% peptone buffer (Becton Dickinson, Sparks, MD). Total aerobic bacteria assays were performed using Aerobic Count Plate Petrifilm™ (3M, Saint Paul, MN). Total coliform and *E. coli* assays were performed using Coliform/ *E. coli* Petrifilm™(3M, Saint Paul, MN). Total *Enterococcus* assays were performed using KF streptococcal agar (Becton Dickinson, Sparks, MD). All data were treated as continuous. To accommodate samples below the limit of detection (LOD), a value of 5 cfu/ml was imputed. The imputed value of 5 cfu/ml is halfway between the LOD (10 cfu/ ml) and 0. Further

information regarding sample collection and processing are detailed in previous studies [15, 41].

Statistical Analysis

Data were analyzed using SAS 9.3 (SAS Institute Inc., Cary, N.C.) at alpha= 0.05. The normal, lognormal, Poisson, gamma, and negative binomial distributions were fitted to the data. This was accomplished by modeling intercept-only models using proc genmod. To model the lognormal distribution, the data was first transformed by taking the natural logarithm of the indicator concentration (performed in SAS 9.3). Of the 13 types of produce collected, 3 types were analyzed individually: cabbage (n=109), cantaloupe (n=42), and cilantro (n=141). These produce types were selected because they had the greatest number of samples. The concentration results from all locations and collection dates were combined. The number of samples for the 13 produce types was as low as 3, and as high as 141. Histograms representing the concentrations were created using Microsoft Excel (2010). A secondary analysis was done for each of the produce-indicator combinations with values below the LOD removed.

The Pearson's Chi-square (χ^2) statistic test was used to test the goodness of fit. It is capable of comparing the observed frequencies of values with the expected frequency from a theoretical distribution [44]. The χ^2 test statistic can be calculated with the following equation:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for bin *i* and E_i is the expected frequency for bin *i*. The degrees of freedom for the Chi-square tests are *n-1*. Expected values were computed with maximum likelihood estimation [45]. A p-value greater than 0.05 indicated that there was no evidence to imply a lack of fit. If the p-value was less than 0.05, it was determined that the distribution did not fit the data. In other words, if the p-value was greater than 0.05, the distribution was considered to fit the data.

Following examination of the χ^2 values, Akaike's Information Criterion (AIC) statistics were compared between distributions that did not indicate a lack of fit, from the Pearson's Chi-square analysis. The AIC is another goodness of fit statistic. The AIC is a likelihood-based statistic that compares the probability that a distribution fits the data (reviewed in [46]). The model with the smallest AIC value was considered to have the best fit [46]. The AIC can be calculated with the following equation:

$$AIC = 2k - 2\ln(L)$$

where *k* is the number of parameters in the model (*k*=1 for the tested models, since no predictors were included), and L is the value of the maximized likelihood function. The final choice for the best model was the model that had a Chi-square p-value greater than 0.05, and the lowest AIC value.

RESULTS

Goal

The goal of this study was to assess the fit of 5 different statistical distributions (normal, lognormal, Poisson, gamma, negative binomial) for the concentrations of 4 indicators (aerobic plate count, coliforms, *E. coli*, *Enterococci spp.*) on cabbage, cantaloupe, and cilantro samples collected from the U.S. side of the United States – Mexico border.

Histograms

Histograms allow a visual assessment of the non-transformed concentration data and are able to quickly provide information on the overall distribution shape, and display skewness characteristics. A total of 12 histograms were produced, one per each indicator (aerobic plate count, coliforms, *Escherichia coli, Enterococcus spp.*) for the 3 produce types (cabbage, cantaloupe, cilantro) (Figures 1, 2, 3). All histograms showed a right skew with the majority of indicator (excluding the aerobic plate count) concentration values below the limit of detection (LOD, 5 cfu/ ml). The right tail consistently had a very low number of extreme concentration values, which are the values of most interest when making risk predictions. It is clearly evident that the data were not normally distributed, and the events of most interest (the highest concentrations) occurred rarely.

The cabbage (n= 109) histograms are shown in Figure 1. The aerobic plate count (a) had values ranging from 9,000 to 29,000,000 cfu/ ml, with very few concentrations between 9,550,000 and 29,000,000 cfu/ ml. The coliform concentrations (b) had values ranging from 5 to 3,000 cfu/ ml, with very sparse data throughout, including 33 out of 109 (30%) samples

below the LOD. The *E. coli* (c) values were highly concentrated below the LOD, with 84 out of 109 (77%) values being reported as 5 cfu/ ml. The remaining 15 concentrations were from 15 to 3,350 cfu/ ml, and were distributed sporadically. The *Enterococcus* (d) data had the most consistent distribution of concentrations, ranging from 5 cfu/ ml to 380,050 cfu/ ml. 6 out of the 109 (6%) samples were below the LOD. However, the data was still heavily right skewed.

The cantaloupe (n= 42) histograms are shown in Figure 2. The aerobic plate count (a) had values ranging from 1,490,000 to 35,000,000 cfu/ ml, with concentrations sporadically distributed throughout. The coliform (b) samples only had 3 out of 42 (7%) concentrations below the LOD, with the remaining values distributed relatively consistently from 20 to 26,300 cfu/ ml, but at low frequencies. The *E. coli* (c) concentrations were mostly aggregated (29 out of 42, 69%) below the LOD, with the few remaining concentrations scattered between 15 and 1,850 cfu/ ml. The *Enterococcus* (d) concentrations were the least skewed of any indicator, amongst all produce types. The distribution displays what appear to be two separate peaks occurring near 5,000 cfu/ ml and 24,000 cfu/ ml, with the most extreme concentration at 69,000 cfu/ ml, and no concentrations below the LOD.

The cilantro (n= 141) histograms are shown in Figure 3. The aerobic plate count (a) had a cluster of values near 3,000,000 cfu/ ml, and then a second peak of values near 2,500,000 cfu/ ml, with the concentration of values slowly decreasing until the maximum concentration at 70,000,000 cfu/ ml. The coliform concentrations (b) had a large portion (32 out of 141, 23%) of concentrations below the LOD, with the remaining concentrations randomly distributed until the highest concentration reported at 30,000 cfu/ ml. The *E. coli* concentrations (c), as with cabbage and cantaloupe, were mostly accumulated below the

LOD (93 out of 141, 66%), with rare concentrations reported between 10 and 10,350 cfu/ ml. The *Enterococcus* concentrations (d) ranged from below the LOD to 265,000 cfu/ ml, but many concentrations (27 out of 141, 19%) were below the LOD, with several concentrations reported from 10 to 265,000 cfu/ ml, and no concentrations reported between 69,000 and 235,000 cfu/ ml.

Distribution Fitting

Fitting different distributions to each of the indicator concentrations was performed to determine which of the five tested distributions was the most appropriate for each of the indicator types, within each of the produce types. The output statistics, including Pearson's chi-square values, Akaike's information criteria (AIC), and the full log likelihood values, can be seen in Tables 1, 2, 3. For 10 of the 12 indicators, one of the 5 distributions was selected as the best fit. If only one distribution fit the data, then that was considered the best fit by default. The two exceptions were the coliform and *Enterococci* indicators in cilantro, which did not fit any of the assessed distributions. The lognormal distribution fit all 10 of the indicators that had at least one distribution that fit the data. It was also consistently the best fit for all of the indicators that had two or more distributions indicating a goodness of fit. The normal and Poisson distributions did not fit for any of the indicators, within any of the produce types. The final distribution selections are summarized in Table 4.

The cabbage results are shown in Table 1. For the aerobic plate count and the *Enterococcus* data, there were 2 distributions that statistically indicated a goodness of fit. The lognormal and gamma distributions fit the aerobic plate count, and the lognormal and negative binomial fit the *Enterococcus* concentration. Therefore the AIC was used to select the

best distribution, which was the lognormal distribution for both indicators. The full log likelihood value was in agreement with the AIC. For both the coliform and *E. coli* concentrations, the lognormal was the only distribution that statistically fit.

The cantaloupe results are shown in Table 2. The aerobic plate count, coliform, and *Enterococcus* indicators all statistically fit the lognormal, gamma, and negative binomial distributions. Using the AIC to select the most appropriate model, the lognormal was preferred. The AIC scores were very close for both the gamma and negative binomial, but not nearly as low as the AIC for the lognormal. The full log likelihood value was in agreement with the AIC. The coliform data only fit the lognormal distribution.

The cilantro results are shown in Table 3. The aerobic plate count data statistically fit the lognormal, gamma, and negative binomial distributions. Again, using the AIC score, the lognormal distribution was selected to be the most appropriate. The full log likelihood value was in agreement with the AIC. The *E. coli* indicator only statistically fit the lognormal distribution. The coliform and *Enterococcus* data did not statistically fit any of the 5 selected distributions. This does not mean that the data does not fit *any* distribution; rather it needs to be examined with distributions not included in this analysis.

As a secondary analysis, values below the LOD were removed for all indicator-produce combinations to determine if the best fitting distributions would be different from the first analysis. The aerobic plate counts did not have concentrations below the LOD; therefore the distribution fitting results were not affected. The results for the coliform indicators varied. Only the lognormal distribution fit the coliform data for cabbage samples. The lognormal, gamma, and negative binomial distributions fit the coliform data for cantaloupe samples. None of the distributions fit the coliform data for cilantro samples. For

the *E. coli* indicator, in all 3 produce types, no distributions were found to fit the data when the values below the LOD were removed. The results for the *Enterococcus* indicator varied. The lognormal and negative binomial distributions fit the cabbage *Enterococcus* samples. The lognormal, gamma, and negative binomial distributions fit the cantaloupe data. However, as with other indicators, none of the distributions fit the *Enterococcus* cilantro samples. When one or more distributions were found to fit an indicator, the AIC score determined that the lognormal distribution consistently had the best fit. In summary, there was no overall effect of removing the values below the LOD, since the lognormal distribution remained the best fitting distribution for produce-indicator combinations with concentration values that fit at least one of the five assessed distributions.

DISCUSSION

The primary goal of this analysis was to assess the fit of 5 different statistical distributions (normal, lognormal, Poisson, gamma, negative binomial) to characterize 4 indicator concentrations (aerobic concentration, coliforms, *Escherichia coli*, *Enterococci spp.*) among cabbage, cantaloupe, and cilantro samples that were collected on the U.S. side of the United States-Mexico border region. This analysis found that, of the 5 assessed distributions, the lognormal distribution consistently fit most (10 out of 12) indicators, while the gamma and negative binomial distributions also fit several of the indicators (6 out of 12). For the indicators where either the gamma and negative binomial distributions fit, in addition to the lognormal, the lognormal was consistently selected as the best fit (based on the AIC statistic).

The lognormal distribution

For 10 of the 12 sets of produce-indicator combinations, the lognormal distribution was found to fit the concentration data based on goodness of fit statistics (Tables 1-3). The coliform and *Enterococcus* indicators in cilantro did not fit any of the assessed distributions. The lognormal likely fit these data because it transformed the shape of the distribution, making it less skewed and more normally distributed. For this study, the data were right skewed, with very few high concentration values, leading to a highly skewed shape with a long right tail (Figures 1-3). By applying a lognormal transformation, the skew shifted towards the center, and the overall shape became more normal. Given that these data result from microbial growth, it was expected that the shape of the distribution would be similar to

other microbial studies, where exponential growth of microorganisms leads to a right skew with rare, high concentration values in the right tail (reviewed in [47]). Highly skewed data, such as microbial concentration data, consequently becomes less skewed after natural log transformation [48]. Further studies on the actual growth of these indicators would provide more insight into the shape of the distribution under different factors influencing the presence and growth of these indicators on produce.

The gamma and negative binomial distributions

The gamma and negative binomial distributions also fit several of the indicators from each of the produce types, namely the aerobic plate counts (Tables 1-3). Both of these distributions work well with over-dispersed data, and both have an additional parameter to assist with accommodating such data [32, 33]. The indicator concentrations (aerobic plate count, coliforms, *E. coli, Enterococci spp*.) were very over dispersed, with concentrations reported as low as 5 cfu/ml, to concentrations exceeding 7 million cfu/ ml (aerobic plate count from cilantro, Figure 3). In addition to the gamma and negative binomial distributions working well with over-dispersed data, they also work with well with skewed data [33]. Since these data were both over-dispersed and highly skewed, it is reasonable to expect these distributions to fit this data. These results were not completely unexpected, since previous literature has identified both the gamma [49] and negative binomial [50] as distributions that can be applied to microbiological data. However, these distributions did not fit all of the indicators. In particular, the gamma and negative binomial did not fit indicators that were *exceptionally* concentrated below or near the LOD (see figures 1-3), with hardly any concentration values in the right tail, such as for *E. coli*. The shape of these particular

indicator concentrations is inconsistent with the gamma and negative binomial distributions, which have a slightly more continually decreasing slope in the right tail (as opposed to a drastic reduction in values). Even though all of the distributions are right skewed, with relatively few values in the right tail, the indicators that fit the gamma and negative binomial distributions still have more information between the peak concentration and highest values compared to the indicators that did not fit these distributions.

The lognormal distribution was consistently the best of the 5 distributions

In all cases where 2 or more of the assessed distributions statistically fit the data (determined by the Pearson's chi-square value), the lognormal was determined to be the best fitting distribution (determined by the AIC statistic). There is a large body of literature discussing the similarities between the lognormal and gamma distributions (reviewed in [51], [52]). However, despite the likeness, the lognormal was still found to be the best fit. Incidences where the two tend to be distinguished from each other are when there is sparse information for high concentration values, and the distributions can become prone to over- or under-estimating the mean and variance [52]. The calculated mean and variance are then used for maximum likelihood techniques, which are factored into the overall goodness of fit tests [52]. It is plausible that for this study, where there was sparse information regarding high concentrations, the lognormal distribution performed better at calculating the mean and variance, and leading to an overall better fit. Whereas the lognormal and gamma are continuous distributions, the negative binomial is a discrete distribution. In situations where the variance is high, the sample size is small, and the overall shape is large, the negative binomial becomes less robust, decreasing the ability for MLE techniques to fit the negative

binomial [51]. As seen in figures 1-3, this data was widely dispersed with a large variance, possibly leading to MLE techniques not being capable of fitting the negative binomial to these particular samples.

Strengths and Limitations

This study had both strengths and limitations. The most notable limitation in this analysis was that only 5 distributions were assessed. It would be worthwhile to examine compound distributions, such as the Poisson-lognormal, to determine if these types of distributions can offer a better fit. Strengths of this study include that both continuous and discrete distributions were compared. The data is treated as concentration values (continuous), but these values arise from plate counts, which are discrete. Therefore, both types of distributions can and should be considered since they both have the potential to reflect the resulting shape of the distribution.

Implications

The findings in this analysis can direct future comparisons of microbial concentration data that are collected from fresh produce, and thereby lead to more accurate risk predictions. A concern in this comparison of distributions was the small sample sizes for some of the sampled produce (i.e. curly parsley, excluded from this analysis due to a small sample size of only 3), which is likely to be a recurring issue for studies examining rare indicators. Once the concentration data is examined to determine a best fitting distribution, researchers will have more confidence in their risk predictions.

Conclusions

In conclusion, this analysis found that the lognormal was consistently the best distribution, compared to the normal, Poisson, gamma, and negative binomial, for the majority of the indicators on cabbage, cantaloupe, and cilantro samples. However, all indicator frequencies indicated a left skew, very few high concentrations, and heterogeneity, suggesting that other distributions should probably be assessed before make any statistical inferences. Not assessing the fit of a distribution can result in an over or underestimation of risk.

References

1. Johnston, L.M., *et al.*., *A field study of the microbiological quality of fresh produce of domestic and Mexican origin.* International Journal of Food Microbiology, 2006. **112**(2): p. 83-95.
2. Hoelzer, K., *et al.*., *Produce Consumption in the United States: An Analysis of Consumption Frequencies, Serving Sizes, Processing Forms, and High-Consuming Population Subgroups for Microbial Risk Assessments.* J Food Prot, 2012. **75**(2): p. 328-340.
3. Barak, J.D. and B.K. Schroeder, *Interrelationships of Food Safety and Plant Pathology: The Life Cycle of Human Pathogens on Plants.* Annu Rev Phytopathol, 2012.
4. Peleg, M., M.D. Normand, and M.G. Corradini, *A Study of the Randomly Fluctuating Microbial Counts in Foods and Water Using the Expanded Fermi Solution as a Model.* Journal of Food Science, 2012. **77**(1): p. R63-R71.
5. CDC, *Ongoing multistate outbreak of Escherichia coli serotype O157 : H7 infections associated with consumption of fresh spinach - United States, September 2006 (Reprinted from MMWR, vol 55, pg 1045-1046, 2006).* Jama-Journal of the American Medical Association, 2006. **296**(18): p. 2195-2196.
6. Lynch, M.F., R.V. Tauxe, and C.W. Hedberg, *The growing burden of foodborne outbreaks due to contaminated fresh produce: risks and opportunities.* Epidemiol Infect, 2009. **137**(3): p. 307-15.
7. Barton Behravesh, C., *et al.*., *2008 Outbreak of Salmonella Saintpaul Infections Associated with Raw Produce.* New England Journal of Medicine, 2011. **364**(10): p. 918-927.
8. Jay, M.T., *et al.*., *Escherichia coli O157 : H7 in feral swine near spinach fields and cattle, central California coast.* Emerging Infectious Diseases, 2007. **13**(12): p. 1908-1911.
9. Beuchat, L.R. and J.H. Ryu, *Produce handling and processing practices.* Emerging Infectious Diseases, 1997. **3**(4): p. 459-465.
10. Ijabadeniyi, O.A., *et al.*., *Irrigation Water as a Potential Preharvest Source of Bacterial Contamination of Vegetables.* Journal of Food Safety, 2011. **31**(4): p. 452-461.
11. Jacobsen, C.S. and T.B. Bech, *Soil survival of Salmonella and transfer to freshwater and fresh produce.* Food Research International, 2012. **45**(2): p. 557-566.
12. Leon-Felix, J., *et al.*., *Norovirus Contamination of Bell Pepper from Handling During Harvesting and Packing.* Food and Environmental Virology, 2010. **2**(4): p. 211-217.
13. De Roever, C., *Microbiological safety evaluations and recommendations on fresh produce.* Food Control, 1998. **9**(6): p. 321-347.
14. Soon, J.M. and R.N. Baines, *Food safety training and evaluation of handwashing intention among fresh produce farm workers.* Food Control, 2012. **23**(2): p. 437-448.
15. Johnston, L.M., *et al.*., *A field study of the microbiological quality of fresh produce of domestic and Mexican origin.* Int J Food Microbiol, 2006. **112**(2): p. 83-95.
16. Ailes, E.C., *et al.*., *Microbial Concentrations on Fresh Produce Are Affected by Postharvest Processing, Importation, and Season.* J Food Prot, 2008. **71**(12): p. 2389-2397.
17. Holvoet, K., *et al.*., *Insight into the prevalence and distribution of microbial contamination to evaluate water management in the fresh produce processing industry.* J Food Prot, 2012. **75**(4): p. 671-81.
18. Kim, S.-R., *et al.*., *Comparison of Sample Preparation Methods for the Recovery of Foodborne Pathogens from Fresh Produce.* J Food Prot, 2012. **75**(7): p. 1213-1218.
19. Administration, U.S.F.a.D. *Analysis and evaluation of preventive control measures for the control and reduction/ elimination of microbial hazards on fresh and fresh-cut produce.* 2001 [cited 2013 February 8]; Available from: http://www.fda.gov/Food/ScienceResearch/ResearchAreas/SafePracticesforFoodProcesses/ucm091260.htm.
20. Kase, J.A., *et al.*., *Microbial quality of bagged baby spinach and romaine lettuce: effects of top versus bottom sampling.* J Food Prot, 2012. **75**(1): p. 132-6.
21. Sutton, S., *MICROBIOLOGY TOPICS: Accuracy of Plate Counts.* Journal of Validation Technology, 2011. **17**(3): p. 42.

22. Hedges, A., *Estimating the precision of serial dilutions and viable bacterial counts.* International Journal of Food Microbiology, 2002. **76**(3): p. 207-214.
23. Field, K.G. and M. Samadpour, *Fecal source tracking, the indicator paradigm, and managing water quality.* Water Research, 2007. **41**(16): p. 3517-3538.
24. Kittigul, L., *et al.*., *Detection and characterization of hepatitis A virus in water samples in Thailand.* Journal of Applied Microbiology, 2006. **100**(6): p. 1318-1323.
25. Busta, F.F., *et al.*., *The Use of Indicators and Surrogate Microorganisms for the Evaluation of Pathogens in Fresh and Fresh-Cut Produce.* Comprehensive Reviews in Food Science and Food Safety, 2003. **2**: p. 179-185.
26. Morton, R.D., *Aerobic Plate Count*, in *Compendium of Methods for The Microbiological Examination of Foods*2001, American Public Health Association.
27. Kornacki, J. and J. Johnson, *Enterobacteriaceae, coliforms, and Escherichia coli as quality and safety indicators.* Compendium of methods for the microbiological examination of foods, 2001. **4**: p. 69-82.
28. Paul, A.H., H.D. Robert, and M.S. Linda, *Enterococci*, in *Compendium of methods for the microbiological examination of foods*2001, American Public Health Association.
29. Harwood, V.J., *et al.*., *Validity of the indicator organism paradigm for pathogen reduction in reclaimed water and public health protection.* Appl Environ Microbiol, 2005. **71**(6): p. 3163-3170.
30. Jongenburger, I., *et al.*., *Modelling homogeneous and heterogeneous microbial contaminations in a powdered food product.* International Journal of Food Microbiology, 2012. **157**(1): p. 35-44.
31. Gonzales-Barron, U. and F. Butler, *A comparison between the discrete Poisson-gamma and Poisson-lognormal distributions to characterise microbial counts in foods.* Food Control, 2011. **22**(8): p. 1279-1286.
32. Jongenburger, I., *et al.*., *Impact of microbial distributions on food safety I. Factors influencing microbial distributions and modelling aspects.* Food Control, 2012. **26**(2): p. 601-609.
33. Jongenburger, I., *et al.*., *Impact of microbial distributions on food safety II. Quantifying impacts on public health and sampling.* Food Control, 2012. **26**(2): p. 546-554.
34. Gonzales-Barron, U. and F. Butler, *Characterisation of within-batch and between-batch variability in microbial counts in foods using Poisson-gamma and Poisson-lognormal regression models.* Food Control, 2011. **22**(8): p. 1268-1278.
35. Finkelstein, M.M. and D.K. Verma, *Exposure estimation in the presence of nondetectable values: Another look.* Aihaj, 2001. **62**(2): p. 195-198.
36. Hewett, P. and G.H. Ganser, *A comparison of several methods for analyzing censored data.* Annals of Occupational Hygiene, 2007. **51**(7): p. 611-632.
37. Kundu, D. and A. Manglick, *Discriminating between the log-normal and gamma distributions.* Journal of Applied Statistical Sciences, 2004.
38. Wiens, B.L., *When log-normal and gamma models give different results: A case study.* American Statistician, 1999. **53**(2): p. 89-93.
39. Olaimat, A.N. and R.A. Holley, *Factors influencing the microbial safety of fresh produce: A review.* Food Microbiology, 2012. **32**(1): p. 1-19.
40. Tauxe, R.V., *et al.*., *Evolving public health approaches to the global challenge of foodborne infections.* International Journal of Food Microbiology, 2010. **139, Supplement**(0): p. S16-S28.
41. Johnston, L.M., *et al.*., *A field study of the microbiological quality of fresh produce.* J Food Prot, 2005. **68**(9): p. 1840-1847.
42. Sant'Ana, A.S., *et al.*., *Prevalence and counts of Salmonella spp. in minimally processed vegetables in São Paulo, Brazil.* Food Microbiology, 2011. **28**(6): p. 1235-1237.
43. Englehardt, J.D., *et al.*., *Methods for assessing long-term mean pathogen count in drinking water and risk management implications.* Journal of Water and Health, 2012. **10**(2): p. 197-208.
44. Eisenhart, C. and P.W. Wilson, *Statistical methods and control in bacteriology.* Bacteriological reviews, 1943. **7**(2): p. 57.
45. *What's New in SAS 9.3.* 2012.

46. Yadav, R.K.P., *et al.*., *Bacterial populations on the leaves of Mediterranean plants: quantitative features and testing of distribution models.* Environmental and Experimental Botany, 2004. **52**(1): p. 63-77.
47. Kilsby, D. and M. Pugh, *The Relevance of the Distribution of Micro- organisms Within Batches of Food to the Control of Microbiological Hazards from Foods.* Journal of Applied Microbiology, 1981. **51**(2): p. 345-354.
48. Limpert, E., W.A. Stahel, and M. Abbt, *Log-normal distributions across the sciences: keys and clues.* BioScience, 2001. **51**(5): p. 341-352.
49. Le Marc, Y., C. Pin, and J. Baranyi, *Methods to determine the growth domain in a multidimensional environmental space.* International Journal of Food Microbiology, 2005. **100**(1–3): p. 3-12.
50. El-Shaarawi, A., S. Esterby, and B. Dutka, *Bacterial density in water determined by Poisson or negative binomial distributions.* Appl Environ Microbiol, 1981. **41**(1): p. 107-116.
51. Zhou, M., *et al.*., *Lognormal and gamma mixed negative binomial regression.* arXiv preprint arXiv:1206.6456, 2012.
52. Cho, H.-K., K.P. Bowman, and G.R. North, *A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the Tropical Rainfall Measuring Mission.* Journal of Applied Meteorology, 2004. **43**(11): p. 1586-1597.

Tables

Indicator	Distribution	Pearson Chi Square (d.f.= 108)	AIC	Full Log(*LL*)
Aerobic Plate Count	Normal	135702 x 10^{10}	3599.98	1797.99
	Lognormal	36.12[a]	192.94[b]	-94.47
	Poisson	505154270.01	325624664.26	-1628112331.10
	Gamma	84.38 [a]	3481.78	-1738.89
	Negative Binomial	163.41	3447.73	-1721.86
Coliforms	Normal	49599422.94	1733.40	-864.70
	Lognormal	88.59 [a]	290.73[b]	-143.37
	Poisson	135837.62	90169.16	-45083.58
	Gamma	372.02	1404.69	-700.35
	Negative Binomial	141.79	1406.54	-701.27
E. coli	Normal	14799947.25	1601.58	-798.79
	Lognormal	44.82 [a]	216.34[b]	-106.17
	Poisson	173089.52	42410.39	-21204.20
	Gamma	2024.33	1030.05	-513.03
	Negative Binomial	635.41	1033.97	-514.99
Enterococcus	Normal	10772632454.60	2319.90	-1157.95
	Lognormal	116.55 [a]	320.63[b]	-158.32
	Poisson	1401121.53	1244349.11	-622173.55
	Gamma	182.24	2098.62	-1047.31
	Negative Binomial	78.88 [a]	2099.01	-1047.50

Table 1
Goodness of fit test statistics for cabbage (n=109)
[a]Represents models that statistically 'pass' Goodness of Fit test
[b]Lowest AIC score, indicating best fit compared to other listed models

Indicator	Distribution	Pearson Chi Square (d.f.= 41)	AIC	Full Log(*LL*)
Aerobic Plate Count	Normal	280.76	1460.20	-728.10
	Lognormal	5.35[a]	36.66[b]	-16.33
	Poisson	266105651.52	244668202.27	-122334100.10
	Gamma	19.30[a]	1443.78	-719.89
	Negative Binomial	44.31[a]	1439.74	-717.87
Coliforms	Normal	2955825103.00	882.10	-439.03
	Lognormal	50.31[a]	130.77[b]	-63.39
	Poisson	372097.21	390408.05	-195203.02
	Gamma	46.84[a]	817.21	-406.53
	Negative Binomial	21.12[a]	817.25	-406.62
E. coli	Normal	6772991.07	626.80	-311.40
	Lognormal	29.10[a]	107.78[b]	-51.89
	Poisson	42808.97	22150.14	-11074.07
	Gamma	270.58	441.82	-218.91
	Negative Binomial	80.12	443.11	-219.55
Enterococcus	Normal	10505440297.56	935.37	-465.68
	Lognormal	8.97[a]	58.36[b]	-27.18
	Poisson	546785.05	531075.35	-265536.16
	Gamma	28.46[a]	915.11	-455.55
	Negative Binomial	36.22[a]	915.11	-455.55

Table 2

Goodness of fit test statistics for cantaloupe (n=42)

[a]Represents models that statistically 'pass' Goodness of Fit test

[b]Lowest AIC score, indicating best fit compared to other listed models

Indicator	Distribution	Pearson Chi Square (d.f.= 140)	AIC	Full Log(*LL*)
Aerobic Plate Count	Normal	1.578×10^{13}	4965.39	-2480.70
	Lognormal	81.33[a]	326.56[b]	-161.28
	Poisson	1760894512.00	1412690848.50	-70635423.20
	Gamma	101.23[a]	4811.77	-2403.88
	Negative Binomial	131.10[a]	4783.18	-2389.59
Coliforms	Normal	2794908248.90	2773.27	-1384.63
	Lognormal	185.43	442.77	-219.38
	Poisson	1468701.78	756555.81	-3738276.91
	Gamma	771.79	2129.74	-1062.87
	Negative Binomial	212.64	2131.34	-1063.67
E. coli	Normal	224678131.21	2417.82	-1206.91
	Lognormal	104.60[a]	362.04[b]	-179.02
	Poisson	696638.07	210083.99	-105041.00
	Gamma	2159.99	1532.01	-764.01
	Negative Binomial	511.13	1535.65	-765.82
Enterococcus	Normal	193434269654.00	3370.70	-1683.35
	Lognormal	255.88	488.17[b]	-242.08
	Poisson	17773188.04	5881799.65	-2940898.83
	Gamma	1633.05	2422.33	-1209.17
	Negative Binomial	343.94	2423.42	-1209.71

Table 3

Goodness of fit test statistics for cilantro (n=141)

[a]Represents models that statistically 'pass' Goodness of Fit test

[b]Lowest AIC score, indicating best fit compared to other listed models

Produce Type	Indicator	Best fitting distribution
Cabbage	Aerobic Plate Count	Lognormal
	Coliforms	Lognormal
	E. coli	Lognormal
	Enterococcus	Lognormal
Cantaloupe	Aerobic Plate Count	Lognormal
	Coliforms	Lognormal
	E. coli	Lognormal
	Enterococcus	None
Cilantro	Aerobic Plate Count	Lognormal
	Coliforms	Lognormal
	E. coli	Lognormal
	Enterococcus	None

Table 4
Summary of best fitting distribution type, according to AIC score, by produce type and indicator

Figures

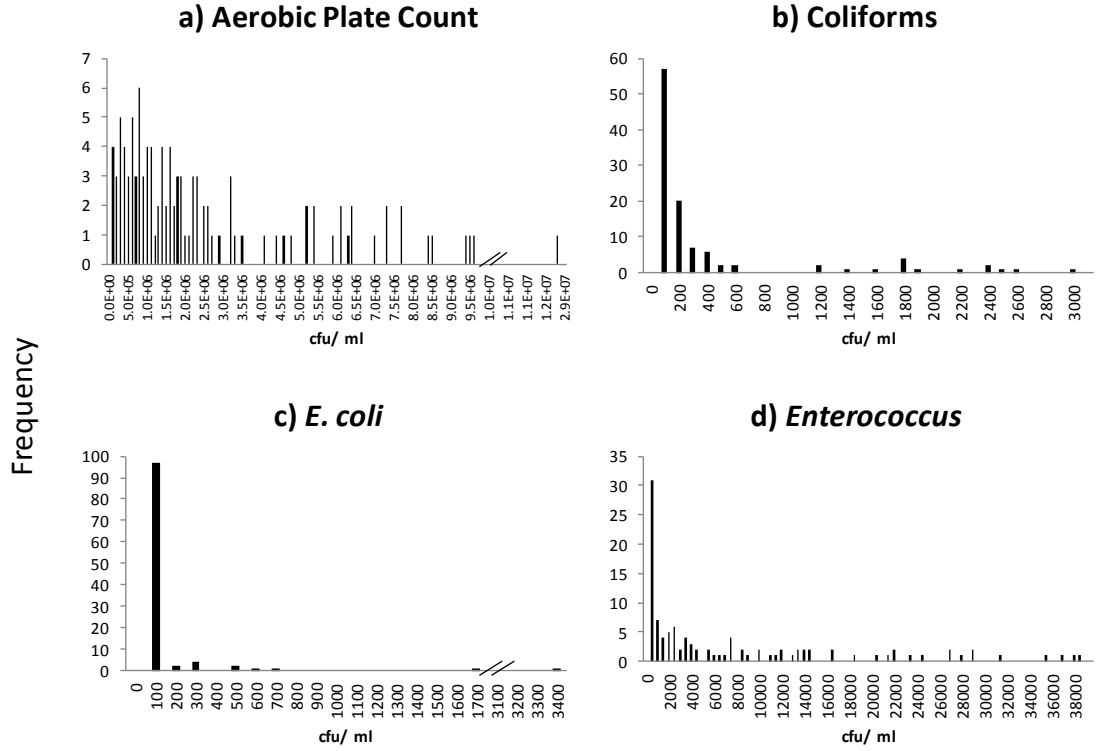

Fig. 1. Histograms of the observed frequencies of CFU counts from cabbage samples (n=109), by a) aerobic plate count, b) coliforms, c) *E. coli*, d) *Enterococcus*. Concentrations below the LOD reported as 5 cfu/ ml.

⫽ = scale break in x-axis

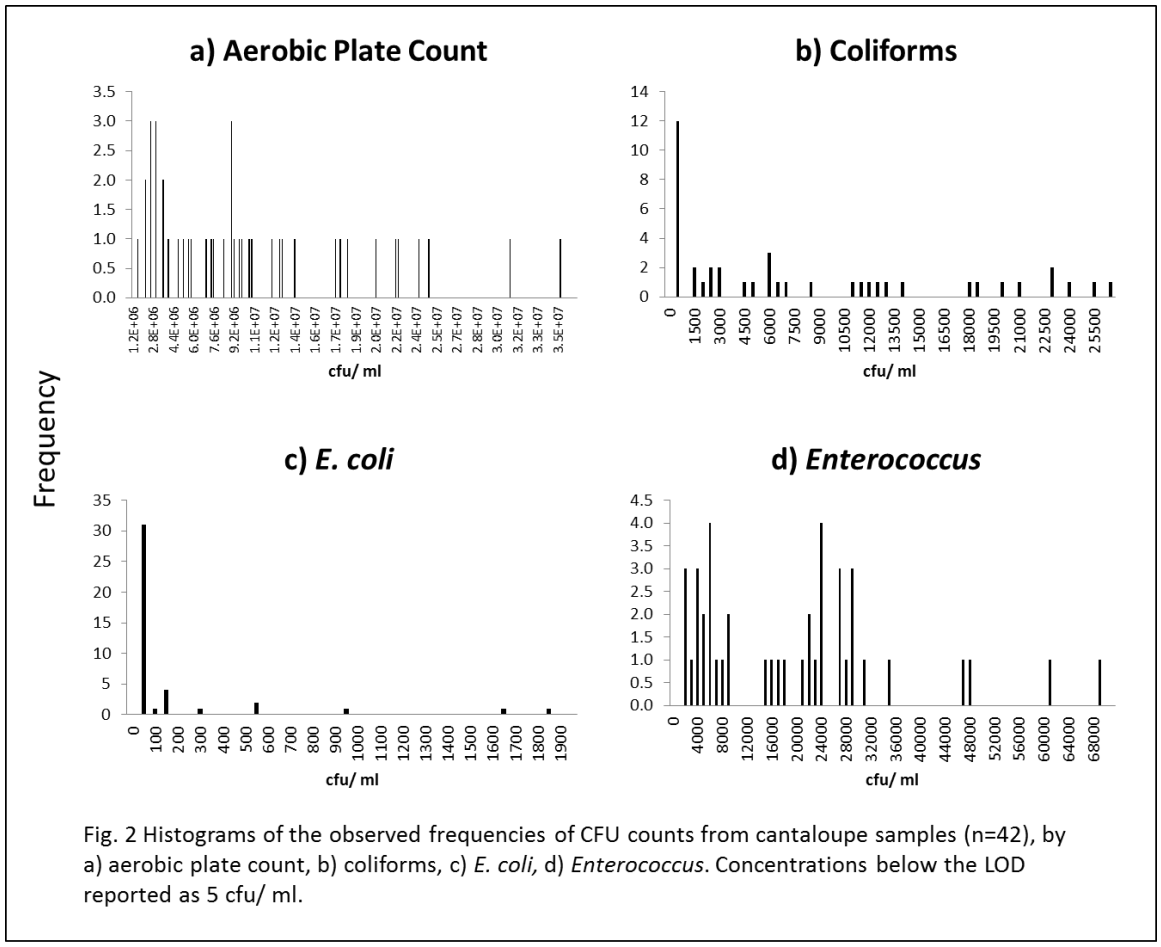

Fig. 2 Histograms of the observed frequencies of CFU counts from cantaloupe samples (n=42), by a) aerobic plate count, b) coliforms, c) *E. coli,* d) *Enterococcus*. Concentrations below the LOD reported as 5 cfu/ ml.

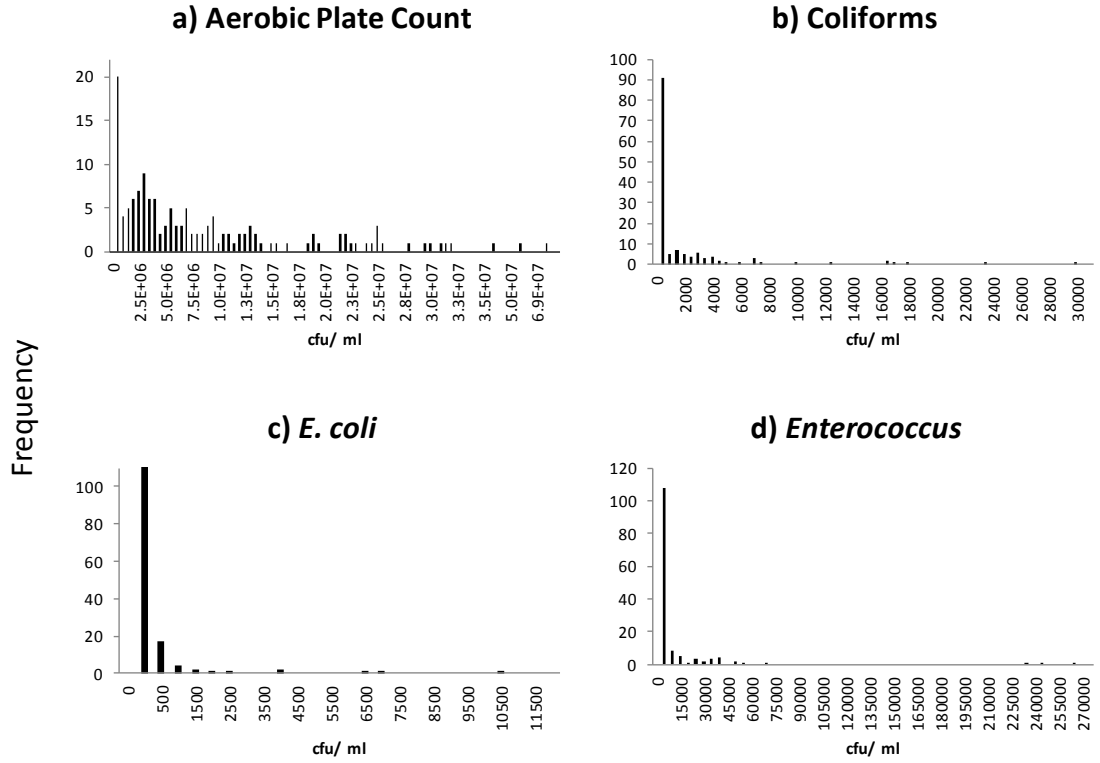

Fig. 3. Histograms of the observed frequencies of CFU counts from cilantro samples (n=141), by a) aerobic plate count, b) coliforms, c) *E. coli,* d) *Enterococcus*. Concentrations below the LOD reported as 5 cfu/ ml.

APPENDIX A: IRB CLEARANCE

EMORY UNIVERSITY | Institutional Review Board

TO: Juan Leon, PhD
Principal Investigator
Global Health

DATE: June 8, 2012

RE: **Continuing Review Expedited Approval**
CR2_IRB00035460
IRB00035460
Identification and Control of Microbiological Hazards in Imported Fresh Fruits and Vegetables: A Field Epidemiological and Intervention Study in Northern Mexico

Thank you for submitting a renewal application for this protocol. The Emory IRB reviewed it by the expedited process on 06/05/2012, per 45 CFR 46.110, the Federal Register expeditable categories F(7) and Subpart D 46.404. This reapproval is effective from **06/29/2012** through **06/28/2013**. Thereafter, continuation of human subjects research activities requires the submission of another renewal application, which must be reviewed and approved by the IRB prior to the expiration date noted above. Please note carefully the following items with respect to this reapproval:

- A waiver of documentation of written/signed informed consent has been renewed.
- A waiver of parental consent has also been renewed.

Documents reviewed with this application:

- Clean Greens scientific protocolCLEAN6-16-10
- consentimiento_enjuaguemanos_07.14.2011
- Informacion-Encuesta Manipulador 23 MAR 2011
- Informacion-Encuesta-Productor-Manager 23 MAR 2011
- Oral Script for Written Consent_FarmManagerSurvey_Spanish_4.26.2011
- Oral Script for Written Consent_FarmManagerSurvey_ver4.26.2011_CLEAN
- OralScript_Hand Rinsing_ver7.14.2011_CLEAN

Any reportable events (e.g., unanticipated problems involving risk to subjects or others, noncompliance, breaches of confidentiality, HIPAA violations, protocol deviations) must be reported to the IRB according to our Policies & Procedures at www.irb.emory.edu, immediately, promptly, or periodically. Be sure to check the reporting guidance and contact us if you have questions. Terms and conditions of sponsors, if any, also apply to reporting.

Before implementing any change to this protocol (including but not limited to sample

size, informed consent, and study design), you must submit an amendment request and secure IRB approval.

In future correspondence about this matter, please refer to the IRB file ID, name of the Principal Investigator, and study title. Thank you.

Sincerely,

Carol Corkran, MPH, CIP
Senior Research Protocol Analyst
This letter has been digitally signed

CC:	Bartz	Faith	Global Health
	Fabiszewski	Anna	Global Health

Emory University
1599 Clifton Road, 5th Floor - Atlanta, Georgia 30322
Tel: 404.712.0720 - Fax: 404.727.1358 - Email: irb@emory.edu - Web: http://www.irb.emory.edu/
An equal opportunity, affirmative action university