Statistical testing for contaminants in an agricultural product

by

Nicholas Mackay

B.S., Southern Utah University, 2020

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Approved by:

Major Professor
Michael Higgins

Copyright

Abstract

Agricultural wholesale products, including different cereal grains, are regularly tested for contaminants such as salmonella. To test for contaminants, individual samples are regularly drawn from the product and these samples are homogeneously mixed to form a single composite sample. A small amount of this composite sample is then selected and tested for the contaminant. Detailed procedures for testing samples for contamination are outlined and regulated by the FDA among other services. Ideally, failure to detect contamination would yield a statistically rigorous limit on the true amount of contamination present in the product.

In this study, we use ideas from risk-limiting auditing and without-replacement sampling to derive a novel test for detecting contamination. We identify a set of conservative, worst-case assumptions that allow us to derive a closed-form probability for failing to detect a contaminant given a pre-specified proportion of contamination present in the product. We then use this probability to develop a risk-limiting statistical test for the null hypothesis that the amount of contamination present is beyond acceptable levels—if no contamination is found, this null is rejected, and our statistical test concludes that the amount of contamination is within a tolerable range. Furthermore, we compute the minimum sample size needed to ensure that, for a pre-specified significance level α, the test rejects the null hypothesis if no contamination is detected. We show that our approach is significantly more powerful than current methods for concluding that an agricultural product is not contaminated. The improvement of our method is especially large for when the amount of agricultural product being tested is small with respect to the size of the individual samples.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Agriculture facilities supply millions of grams of wholesale product, such as cereal grains, to consumers. To ensure that the product is safe for consumption, the product is tested for microbial contaminants like salmonella. The testing process for an agricultural product is as follows. First, samples of the product are repeatedly drawn. Then, these samples are combined into a single composite sample. Finally, this composite sample is then homogeneously mixed, and a small portion of this composite sample is tested for the contaminant. If any of the individual samples contain a detectable amount of contaminant, the composite sample will also test positive—or in other words, a composite sample will only test negative if all individual samples are free from contamination.

Ideally, if no contaminated samples are selected, this should indicate strong evidence that the product is free of contamination. However, if contamination is present in the product, it may be allocated within in the product in such a way to avoid detection by this testing method. That is, the only way to be 100% certain that no contamination is present within the product is to sample the entire product. Hence, there is a need for methods that obtain *statistical* guarantees and estimates on the amount of contamination present in the product.

We consider a specific instance of this problem where we are testing whether a single container of agricultural product is free of contamination. Using ideas from financial and

election auditing (Aslam et al., 2008; Stark, 2008), we develop a conservative, risk-limiting hypothesis test for detecting contamination in this container in this container. This test provides statistical assurance that, if no contaminated samples are selected, the amount of contamination within the container is within acceptable levels (Fienberg et al., 1977). We then demonstrate how to find the requisite sample sizes for this test at any given level of significance α (Wendell and Schmee, 1996).

We show that our approach is more powerful than the approach in Jarvis (2007)—which is currently the best-performing method for this particular application. The gain in power is driven by viewing the sampling procedure as being performed "without replacement" rather than "with replacement," and is largest when container sizes are small (with respect to the size of the sample) (Fienberg et al., 1977).

Note, sampling preparation and testing must follow strict requirements by the FDA and outlined in the bacteriological analytical manual (BAM) (Andrews et al., 2018). Some procedures are discussed as needed; for more details, refer to BAM on how to prepare samples of different classifications of food categories. This study analyzes quality control for the wholesale product and not the actual testing procedures that are performed by lab professionals.

1.1 Background

Agricultural industries undergo required testing for contamination of their wholesale product. To comply with legislative requirements—for example, European Regulation (EC) No 2073/2005 (Commission, 2005)—a "sufficient" number of samples must be drawn from the agricultural product and each sample must be deemed free from contamination before the product is considered safe for consumption. Sample sizes and testing procedures are determined by food category, and are outlined in the Bacteriological Analytical Manual (BAM) (Andrews et al., 2018). Of note, the sampling procedure is required to produce a

"representative" sample of the product. Refer to and Chapters 1 and 5 of Andrews et al. (2018) for more details.

In practice, the sampling procedure is as follows. Facilities mill dehydrated forage products into bulk bags containing kale, alfalfa and other cereal grass that are prepared for wholesale. While product is being milled, an automated sampler repeatedly draws samples of the product. These samples are drawn throughout the entire milling process. Then, these samples are combined and homogeneously mixed, forming a single composite sample. A portion of the composite sample is then tested for the contaminant. Compositing the batches and mixing homogeneously is used to provide an accurate representation of the finished product. This sampling process is called *composite sampling*.

Composite sampling is common when preparing samples for testing. This method will often lead to a sample that is representative the population while being more cost effective than testing individual batches (Jarvis, 2007; Patil, 2002). If any individual sample used to form the composite sample is positive for contamination, and the test for the contaminant is "sufficiently sensitive," then the portion tested from the composite sample test positive. Or in other words, the test from the composite sample will only be deemed contamination free if all individual samples are contamination free. Therefore, testing for contaminants from the composite sample will allow for conclusions about the contamination of the entire wholesale product. The major drawback of the compositing sampling method arises from the sensitivity of the salmonella test. If there is a lack of accuracy to detect a small amount of contamination in large quantity of product, then composite sampling might not be the correct approach.

Another drawback of the current methodology is that it does not provide a method for estimating the percent of contamination upon a positive test result (Jarvis, 2007). The level of contamination is instead treated as an unknown parameter and various levels of contamination are hypothesized to determine requisite sample sizes for testing. A variety of assumed contamination levels are analyzed. The more contamination there is, the easier it

should be to detect contamination.

Jarvis (2007) uses the binomial distribution for computation of sample sizes. Because of the binomial outcome being a positive or negative instance, it is important to recognize and understand that even if negative tests of high batch size is the result, it should never be concluded or assumed that the product is "contamination free" (Jarvis, 2007). The statistical probability of accepting the product is largely based on the prevalence of contamination.

1.2 Literature Review/Previous work

The problem of testing for a contaminant is closely related to problems often studied in financial and election auditing. We are repeatedly drawing samples from a population. We conclude that the population does not contain "significant contamination" only if all samples are free from contamination. If any "contaminated" samples are drawn, strong action is taken—for example, for an agricultural product, the product is destroyed if it tests positive for the contaminant.

In financial auditing, financial records are sampled and inspected to detect "material overstatement". (Fienberg et al., 1977; Miratrix and Stark, 2009) Each sampled record is inspected and the book value of each record is determined. If the reported value of any sampled record is significantly larger than its book value—or, in the context of testing for contamination, if the financial record is "contaminated with overstatement"—a full audit of all financial records is performed (Fienberg et al., 1977). If the overstatement of all sampled records are within a given level of tolerance, the audit concludes, and the auditor concludes there is no material overstatement.

In election auditing, an election result is legally defined as the result obtained after a full hand count of ballots. (Miratrix and Stark, 2009; Stark, 2008) However, to reduce the time and manpower required to complete an election, ballots are often counted by machines (Higgins et al., 2011). Discrepancies between machine counts hand counts may occur—in many

elections, small, inconsequential differences between these counts are common (Aslam et al., 2008; Stark, 2008). Causes of these discrepancies may include human error in filling out a ballot (e.g. using a check mark instead of filling in a bubble to denote a preference), software bugs, and malicious attacks (Stark, 2008). An audit may be performed to ensure that the machine-count outcome does not differ from the result a full hand count would show. A set of precincts is sampled, and ballots within these precincts are hand counted (Aslam et al., 2008). If a large discrepancy between the hand count and machine count within a precinct is observed—that is, if a precinct is "contaminated" with a discrepancy between the machine count and the hand count—a full hand count may be triggered (Stark, 2008). The thoroughness of this auditing procedure depends on the margin of victory—more precincts need to be sampled to verify the election result if the margin of victory is small (Aslam et al., 2008; Lindeman and Stark, 2012; Miratrix and Stark, 2009; Stark, 2008). In large elections, a margin of victory of only a few hundred votes will often trigger a full hand count automatically.

If units are uniform and sampled with replacement, testing for the presence or absence of a characteristic triggers a binomial solution. The work by Jarvis (2007) mentioned above examines this methodology combined with composite sampling. "Any result that can be classified on the basis of a presence or absence is governed by the binomial distribution. Typically, the binomial considers n number of independent trials with replacement. This is similar to a Poisson distribution that can provide a probability of a number of independent events occurring in a time interval, volume, area and distance. Comparing the two probabilistic solutions using composite sampling, Jarvis showed "that when testing [for single salmonella] is 100% effective, the probability of detection just depends on the total quantity of sample tested, not the number of individual tests" (Jarvis, 2007). He concluded the method requires sufficient sensitivity to detect salmonella in large quantities. If the sensitivity is not reliable, than certainty does not exist and test results could indicate false-negatives.

When units differ in size, the sampling scheme and/or methods used to analyze the sam-

ple may be altered to reduce the cost of a full audit (Aslam et al., 2008). For example, when the maximum discrepancy (error) for a unit can be bounded prior to sampling, the sampling scheme can be altered to increase the likelihood of selecting units with a large potential for error—for example, sampling with probability proportional to the size of this error bound (PPEBWR sampling) Aslam et al. (2008). Additionally, the method for analyzing the sample may use information about the magnitude of the discrepancy found within each sampled unit to obtain a more precise estimate on the total discrepancy. The trinomial bound audit (Miratrix and Stark, 2009) provides an extension to the binomial approach—units are sampled using PPEBWR sampling, and discrepancies are classified to be either 0, small, or large. Similarly, Fienberg et al. (1977) use PPEBWR sampling and the multinomial distribution to estimate the total amount of misstatement, though this method may be computationally infeasible when many small, but non-zero, discrepancies are found.

A fundamental component to this application is the concept of a risk limiting audit (Aslam et al., 2008; Higgins et al., 2011; Lindeman and Stark, 2012; Stark, 2008), which is a common approach especially in election auditing. In short, a risk limiting audit method inverts the typical hypotheses for a test for detecting discrepancies. In the context of election auditing, the null hypothesis is that there is sufficient misstatement in an election to overturn the result and the alternative hypothesis is that the amount of misstatement will not overturn the election (Fienberg et al., 1977; Higgins et al., 2011; Lindeman and Stark, 2012). Hence, if we reject the null hypothesis, we have a statistical (not absolute) confirmation that initial election result is correct. If the null hypothesis is not rejected, further action can be taken to obtain the correct result—for example, performing a full hand count.

Risk-limiting audits may be performed as a set of sequential tests as well (Lindeman and Stark, 2012; Stark, 2008). An initial audit of sampled units is performed, and if not enough evidence is present to reject the null hypothesis, further units may be sampled and audited. This process can then be repeated until either sufficient evidence against the null hypothesis is obtained or a full audit is performed. In the latter case, we are guaranteed

to either confirm or reject the null hypothesis with complete certainty. In the context of election auditing, this risk limiting procedure suggests an "intelligent incremental recount that stops when the audit provides sufficiently strong evidence that a full hand count would confirm the outcome" Lindeman and Stark (2012).

1.3 Outline

The rest of the report is organized as follows. In Chapter 2, our method for quality control is described in detail. In Section 2.1 we describe the notation and framework of our risk-limiting hypothesis test. We outline conservative assumptions that allow us to use the hypergeometric distribution to compute p-values for this test. We then describe how to find the requisite sample sizes necessary to reject the null hypothesis for any significance level α if no contaminated samples are selected. We apply our methodology under various scenarios in Section 2.2, and make comparisons against the method in Jarvis (2007), which is derived assuming with-replacement sampling. We discuss results from this simulation study in Section 2.3. Chapter 3 concludes.

Chapter 2

A Risk-Limiting Test For Detecting Contaminants

A bulk container of agricultural product is required to be tested for contamination. A sampler will draw samples of constant sizes without replacement from this container and homogeneously mix the batches into one composite sample. A small selection from that composite sample is then tested for contamination. If any individual sample is contaminated, then the selection from the composite sample will test positive for the contaminant. The question is how to determine the number of batches that need to be collected and how accurate is the test result.

2.1 Notation and Preliminaries

Suppose that the container contains S grams of product, and each sample contains s grams of product. Let N denote the total number of samples that can be drawn from the container; that is, $N = \lceil S/s \rceil$. Let n denote the true number of samples drawn from the wholesale product. Once n samples are drawn, a composite sample comprised of ns grams is created and homogeneously mixed, and a small selection from this composite sample is tested for salmonella.

If the selection tests positive, then the container is considered *contaminated*. If it tests negative, the container is considered *not contaminated*. However, since the entire container is not sampled, there may still be some contamination present in the container; the entire container will need to be sampled and tested in order to guarantee that it is completely free of the contaminant. Instead, we aim to construct a procedure such that, given a pre-specified acceptable contamination percentage γ, we can determine the sample size n necessary to ensure that a contamination level greater than γ is guaranteed to be detected with a given confidence level $1 - \alpha$.

2.1.1 Statistical testing of contamination

Detecting contamination can be posited as a "risk-limiting" hypothesis test. Typically, when testing for the presence of a contaminant, the hypothesis test is constructed such that the null hypothesis (H_0) is that the contaminant is not present and the alternative hypothesis (H_a) is that the contaminant is actually present. However, failing to reject H_0 is not the same as concluding H_0 is correct. For example, when testing for salmonella, failing to accept an alternative hypothesis that there is not a statistically significant amount of salmonella contamination is not the same as a statement that the amount of salmonella is statistically likely to be below some pre-specified threshold.

Instead, risk-limiting hypotheses tests switch these hypotheses—that is, H_0 is that there is a material presence of the contaminant and H_a is that there is not a material presence. That is, the risk-limiting hypothesis testing paradigm is the appropriate approach for testing for the **absence** of the contaminant.

Under the risk-limiting approach, the null and alternative hypotheses are constructed as

$$H_0 : p \geq \gamma \quad \text{(contaminated)},$$
$$H_a : p < \gamma \quad \text{(not contaminated)},$$

where p is the true proportion of contamination in the container. For this approach, we will reject the null for the alternative (that is, conclude no contamination) if no samples are contaminated. If any sample is contaminated, we will conclude that the entire container is contaminated and unsafe for consumption.

We can ensure certain confidence for our test by pre-specifying the Type I Error. However, computing a p-value—in this case, the probability of not detecting contamination when at least γ is present—can be challenging. This will be a function of the number of samples N contained in a contained, the sample size n, and the amount of contamination under the null hypothesis γ. Additionally, this will depend on the true distribution of the contamination within the container; there are innumerable ways that γ contamination could exist within the container.

To make the quality control problem tractable, we make some simplifying assumptions. These assumptions are designed to control the probability of falsely concluding a container to be free of the contaminant even if the allocation of the contaminant is "as hard to detect as possible."

2.1.2 Sampling framework

Let us assume that the product is divided into uniform, pre-defined, separated batches of, at most, s grams. Each batch is either contaminated or not contaminated (a partially contaminated batch is considered to be contaminated). For a given acceptable contamination level γ, we first determine the minimum number of batches that can contain this contamination. As we will see, the probability of detecting contamination is smallest when contamination is isolated to as few batches as possible. This occurs when $N\gamma$ batches are completely contaminated and the remaining batches are free from contamination.

We then assume that we draw a simple random sample of n batches from the finite number of batches N. In practice, a sample drawn in practice could be comprised of multiple of our pre-defined batches that comprise the product. However, in this case, the detection becomes

easier (i.e. we may be able to sample from multiple batches with one draw) than if we were limited to the pre-defined batches as assumed. Determining n to ensure accurate detection of the contamination is a primary goal of this study.

We note that the sampling of the product is often done systematically rather than at random. If contamination only occurs at regular intervals within the container, the assumption of simple random sampling will be unreasonable.

2.1.3 Testing using the Hypergeometric Distribution

Suppose we have a set of N units, with K of these units having some characteristic of interest. Suppose further that n units are sampled from the N units at random without replacement. Then, the number of units X in the sample that have the characteristic follows a hypergeometric distribution:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{2.1}$$

We will use this distribution to compute p-values for our risk-limiting hypothesis test. This test will be more powerful test than those that rely on the binomial distribution, which inherently assumes that samples are drawn with replacement.

We apply this distribution to our problem of detecting contamination (contamination is a success). Recall that N is total number of batches in the container and γ is the hypothesized proportion of contaminated batches. That is, γ contamination can be contained in no fewer than $K = \lceil N\gamma \rceil$ batches. We only conclude that a product is safe for consumption (i.e. has less than γ contamination) if we do not sample any contaminated batches: $k = 0$. Thus, the probability of concluding that a product is safe for consumption when at least γ contamination is present is no larger than

$$P(X = 0) = \frac{\binom{N-\lceil N\gamma \rceil}{n}}{\binom{N}{n}}. \tag{2.2}$$

Since we only reject the null hypothesis $H_0 : p \geq \gamma$ if no contaminated samples are selected, this becomes our p-value for our risk-limiting hypothesis test. If this p-value is less than the significance level α, we conclude that the product is safe for consumption.

Theoretically, it is possible to generalize this test to allow for leniency in the number of contaminated samples—that is, to allow the test to reject the null hypothesis even if one or more batches drawn are contaminated. In this case, the test statistic would be the number of positive samples k and the p-value would be the probability of drawing k or fewer positive samples. However, in practice, this is not feasible. Samples are homogeneously mixed together to form a single composite sample, and a selection from the single composite sample is tested for contamination. If this test comes back positive, it is impossible to determine how many contaminated individual samples contributed to this positive test.

2.1.4 Choosing the sample size

It is critical to ensure that the sample size n is large enough to ensure that the p-value is below the α threshold when no contaminated samples are selected; otherwise, under this framework, it would be impossible to conclude that a product is safe for consumption. However, larger sample sizes also lead to increased costs and less usable product after testing for contamination. Thus, for a given significant level α, we want to select the smallest sample size possible such that the p-value is less than α if no samples are contaminated.

To find this sample size n, we compute the probability (2.2) across all possible values of the sample size n. We are then able to identify the smallest n that obtains a probability less than α.

2.2 Simulation and Results

We now determine the requisite sample sizes n under a variety of scenarios. We specify for each scenario the number of batches N within a product, the significance level α, and the

contamination level γ. We then find the smallest value of n such that the probability (2.2) is less than α. For this study, we consider the number of batches $N = [30, 50, 100, 1500]$, hypothesized contamination levels $\gamma = [2.5\%, 5\%, 7.5\%, 10\%, 15\%, 20\%]$, and significance levels $\alpha = [0.1, 0.05, 0.01]$, and note that this approach can be used for any other choice of parameters. Results are found in Tables 2.1–2.5. Figures 2.1–2.4 compare our results to those using the method in Jarvis (2007) which assumes with-replacement sampling.

To demonstrate how a practitioner would use this method, suppose a wholesale container contains a total of $S = 5500g$, and an automated sampler consistently draws $55g$ per extraction. Thus, $N = [S/s] = 5500g/55g = 100$. Suppose further that regulations require that the testing procedure identifies a container with a $\gamma = 10\%$ contamination percentage with at least 90% accuracy (hence, $\alpha = 0.1$). From Table 2.4, this will require the practitioner to draw 20 samples of product.

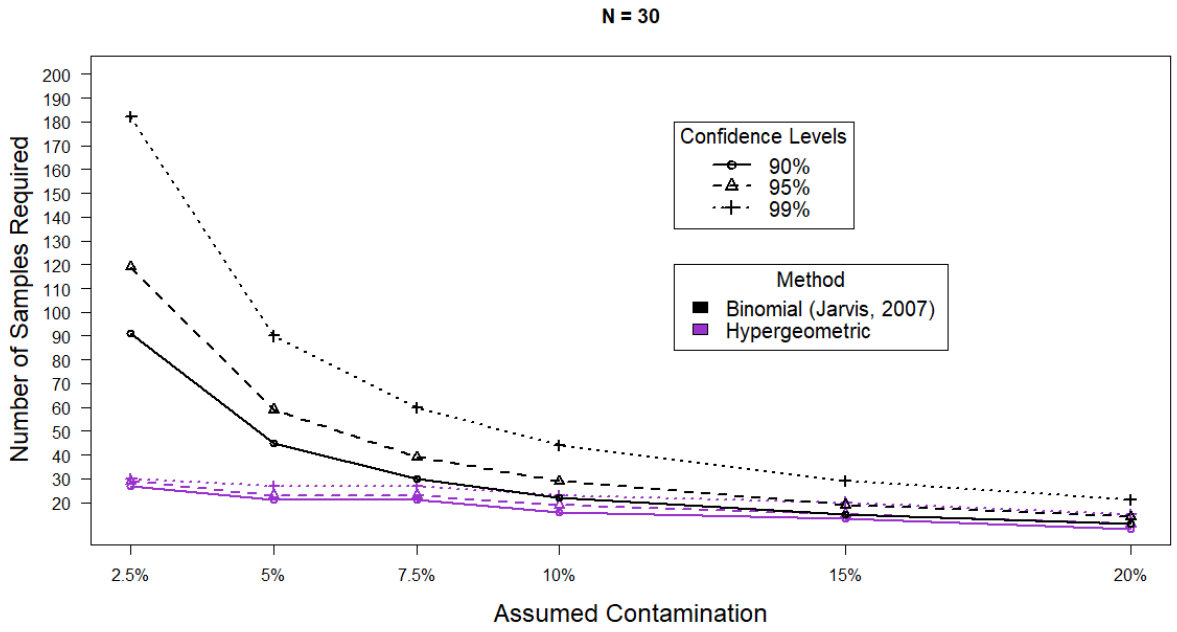

Figure 2.1: *Comparing our approach to the binomial approach when $N = 30$. This graph gives the minimum number of samples n required to conclude that a container is contamination-free if all samples are clean. Sample sizes are computed for varying thresholds of the contamination level γ*

Table 2.1: *Hypergeometric required n for each γ when N=30.*

γ *Contamination*	2.5%	5%	7.5%	10%	15%	20%
90% Confidence	27	21	21	16	13	9
95% Confidence	29	23	23	19	15	11
99% Confidence	30	27	27	23	20	15

Table 2.2: *Minimum samples n to conclude contamination-free using a binomial approach.*

γ *Contamination*	2.5%	5%	7.5%	10%	15%	20%
90% Confidence	91	45	30	22	15	11
95% Confidence	119	55	39	29	19	14
99% Confidence	182	90	60	44	19	21

2.3 Discussion

Note that the binomial probability is not affected by the number of batches N contained in the container of product; the probability of detecting no contaminated batches if γ contamination is present is

$$(1-\gamma)^n.$$

Thus, the requisite sample size n for the different simulations will remain constant across differing number of batches N (see Table 2.2). Additionally, recall that the binomial approach examines independent samples with replacement which accounts for the possibility of drawing the same sample multiple times. This can lead to inefficient samples and increase the required n. Therefore when examining the results, we see sample sizes needed that exceed that N.

When $N = 30$, (see Figure 2.1) our hypergeometric approach leads a dramatic difference in the number samples that need to be drawn, most noticeably when γ is small. When larger amounts of contamination are present, detection becomes easier and thus results in smaller n for the hyper geometric and about 4.5 times smaller n for the binomial results. With 5% contamination, our method shows $n = 21, 23, 27$ for confidence $90\%, 95\%, 99\%$ respectively. If we compare that to the binomial results of $n = 45, 59, 90$ with the same confidence, at our worst, the samples needed is cut in half and can result in up to more than 3 times

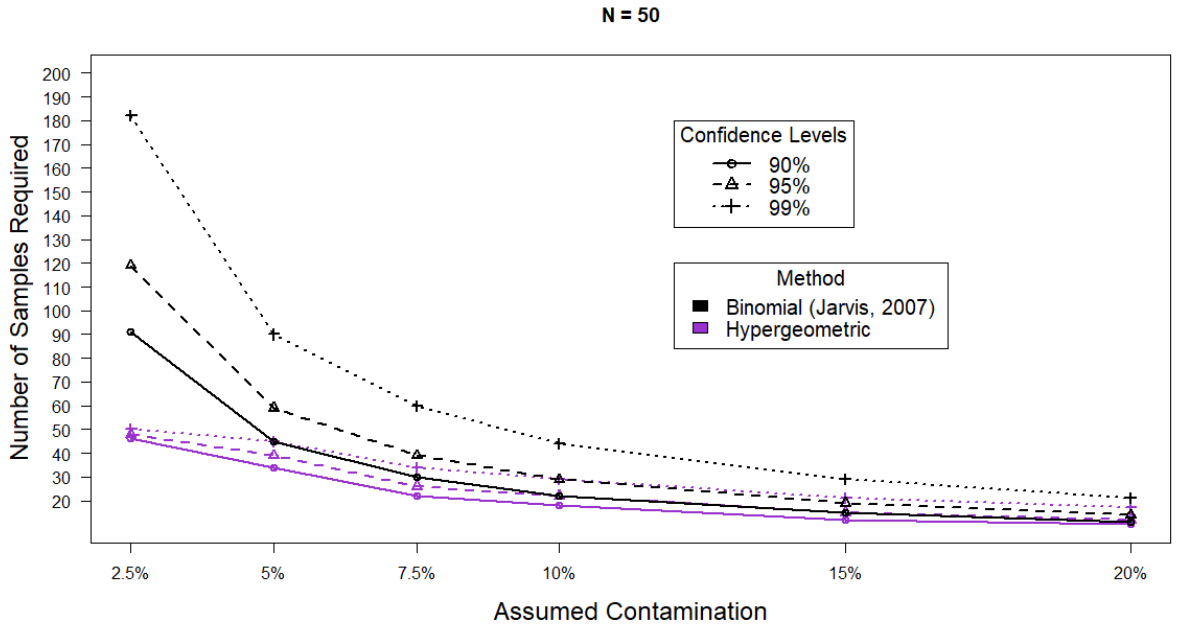

Figure 2.2: *Comparing our approach to the binomial approach when $N = 50$. This graph gives the minimum number of samples n required to conclude that a container is contamination-free if all samples are clean. Sample sizes are computed for varying thresholds of the contamination level γ.*

Table 2.3: *Results for required n for each γ when N=50.*

γ *Contamination*	2.5%	5%	7.5%	10%	15%	20%
90% Confidence	46	34	22	18	12	10
95% Confidence	48	39	26	22	15	12
99% Confidence	50	45	34	29	21	17

more efficient. The hypergeometric approach becomes even more efficient when considering $\gamma = 2.5\%$.

Again, consider $\gamma = 5\%$ and compare Tables 2.3, 2.4, 2.5 ($N = 50$, $N = 100$, and $N = 1500$). The results show a relatively small increase in the number of samples required when doubling the batch size. Even more so when increasing N by 30 times.

As N increases, we see that the required n for the binomial and hypergeometric approach get closer. This is expected—in general, simple random sampling can be approximated by sampling with replacement when population sizes are large (Lohr, 2021)—and this trend is

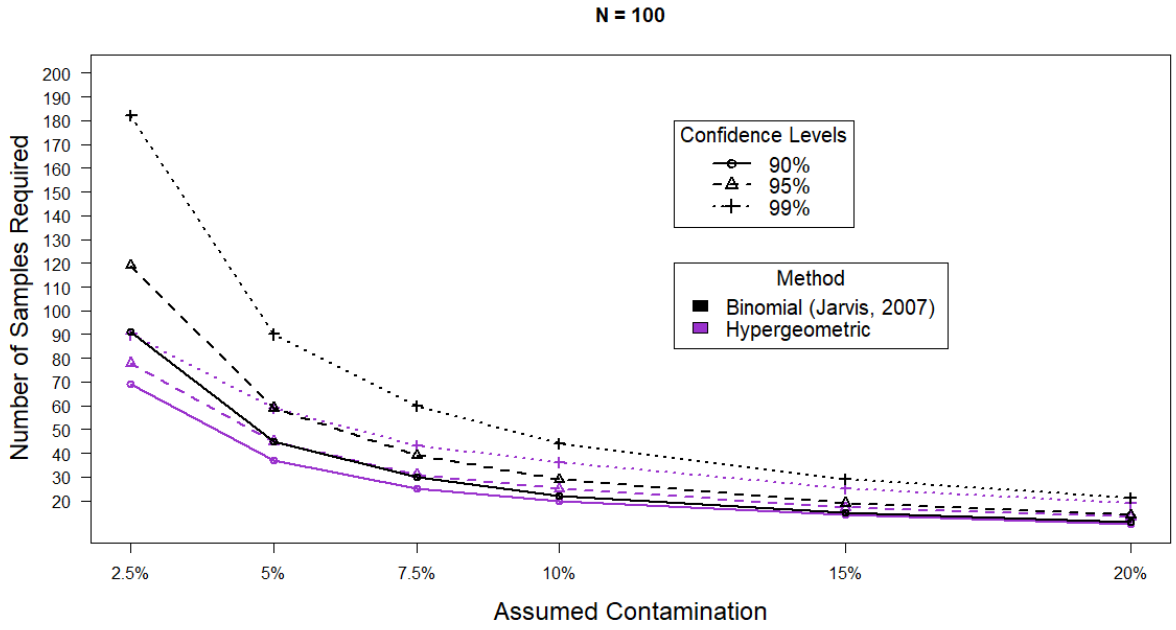

Figure 2.3: *Comparing our approach to the binomial approach when $N = 100$. This graph gives the minimum number of samples n required to conclude that a container is contamination-free if all samples are clean. Sample sizes are computed for varying thresholds of the contamination level γ.*

Table 2.4: *Results for required n for each γ when N=100.*

γ *Contamination*	2.5%	5%	7.5%	10%	15%	20%
90% Confidence	69	37	25	20	14	10
95% Confidence	78	45	31	25	17	13
99% Confidence	90	59	43	36	25	19

shown through Figures 2.1, 2.2, 2.3, and 2.4. With low contamination and a high desired confidence level, it becomes necessary to sample almost, if not all, the entire product. In reality, 99% confidence is not obtainable for a wholesale company since you'd need to examine the entire product. Therefore the pre-determined confidence levels should be carefully chosen based on how much product needed for wholesale and if it is affordable to sample large amounts.

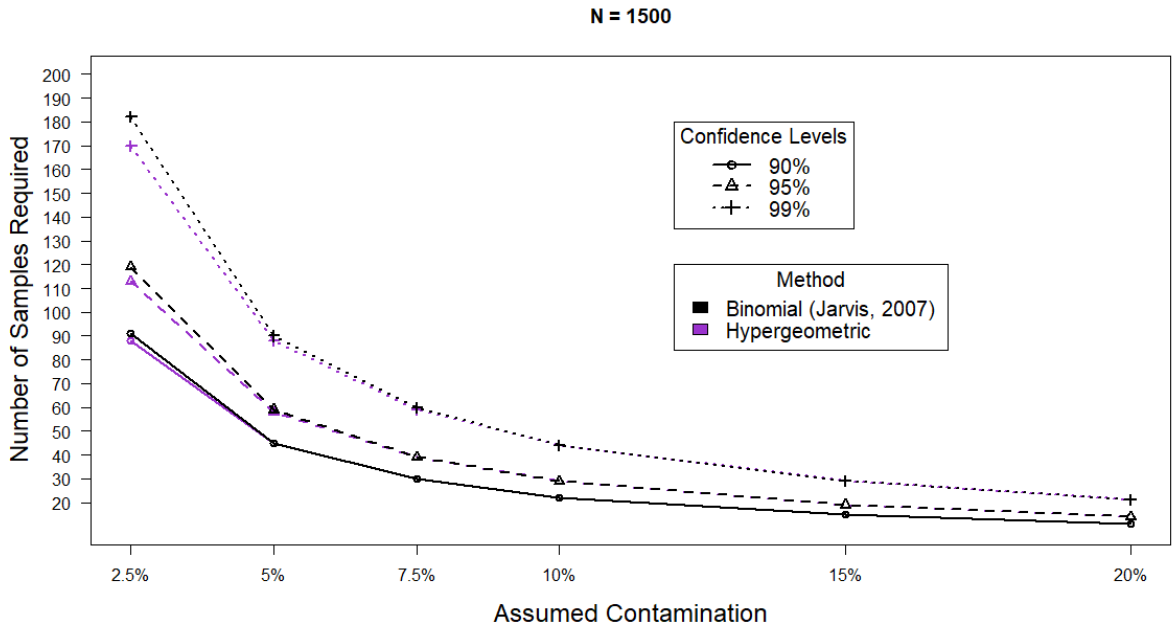

Figure 2.4: *Comparing our approach to the binomial approach when* $N = 1500$. *This graph gives the minimum number of samples* n *required to conclude that a container is contamination-free if all samples are clean. Sample sizes are computed for varying thresholds of the contamination level* γ.

Table 2.5: *Results for required* n *for each* γ *when N=1500.*

γ *Contamination*	2.5%	5%	7.5%	10%	15%	20%
90% Confidence	88	45	30	22	15	11
95% Confidence	113	58	39	29	19	13
99% Confidence	170	88	59	44	29	21

Chapter 3

Conclusion

Accurate detection of contamination in an agricultural wholesale product is essential. To detect contamination, samples of the product are repeatedly drawn, and these samples are combined into a single composite sample. This composite sample is then homogeneously mixed, and a portion of this composite sample is tested for the contaminant. If any individual sample is contaminated, the portion from the composite sample will test positive for the contaminant. Sampling procedures must follow strict guidelines outlined by the FDA, and are outlined in the bacteriological analytical manual (Andrews et al., 2018).

We borrow ideas from risk-limiting auditing to devise a statistical test for detecting contamination (Aslam et al., 2008; Fienberg et al., 1977; Lindeman and Stark, 2012; Stark, 2008). The null hypothesis for our test assumes unsafe levels of contamination in the product, and by rejecting this null, we conclude that the amount of contamination is low enough that the product is safe for consumption (Higgins et al., 2011; Stark, 2008). This approach requires some simplifying assumptions. Conceptually, we assume the sample is separated into batches where we sample without replacement. The prevalent contamination level is pre-determined and impacts the difficulty of detecting an contamination in these batches. We then can use the hypergeometric distribution to compute the probability of not sampling a contaminated batch—a p-value for this test. Note, this probability assumes that the

allocation of contamination (γ) throughout these batches makes detection of contamination as hard to detect as possible. Therefore, if our test concludes that the product is free from contamination at a significance level α, the true significance level at which we would reject the null hypothesis is, in practice, much less than α.

We perform a simulation study to assess the efficacy of our method compared to previous methods that rely on with-replacement sampling theory (Jarvis, 2007). We show that our approach provides a significant improvement to current methods. This improvement is largest when the number of batches within the product is small.

3.1 Future Work

For future work, we may consider methods for better estimating and assessing how contamination is allocated within a product. Our method used simplifying assumptions that consider the most difficult allocation of contamination. Less conservative assumptions, such as limiting the amount of contamination allowed in a sampled batch, may dramatically increase the power of our method, and may be reasonable to make in some instances. Additionally, our hypothesis test may be inverted to devise a (very conservative) upper confidence bound on the amount of contamination present in a product.

Statistical modeling may also help improve in detecting contamination. Most of the time, testing for contamination yields a 'true' or 'false' result and not an actual measurement. A potential method that would combat this nuisance may involve using Bayesian techniques to model the prevalence of contamination (Meeden, 2003). The contamination level is then represented as a distribution and we could consider more realistic values based on a small amount data.

Bibliography

Wallace H Andrews, A Jacobson, and Thomas Hammack. Bacteriological analytical manual (bam) chapter 5: Salmonella. *Bacteriological Analytical Manual*, 110, 2018.

Javed A. Aslam, Raluca A. Popa, and Ronald L. Rivest. On auditing elections when precincts have different sizes. In *USENIX Workshop on Accurate Electronic Voting Technology*, 2008.

European Commission. Commission regulation (ec) no 2073/2005 of 15 november 2005 on microbiological criteria for foodstuffs. *Off. J. Eur. Union*, 50:1–26, 2005.

Stephen E Fienberg, John Neter, and Robert A Leitch. Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, 72(358): 295–302, 1977.

Michael J Higgins, Ronald L Rivest, and Philip B Stark. Sharper p-values for stratified election audits. *Statistics, Politics, and Policy*, 2(1), 2011.

B. Jarvis. On the compositing of samples for qualitative microbiological testing. *Letters in Applied Microbiology*, 45(6):592–598, 2007. doi: https://doi.org/10.1111/j.1472-765X.2007.02237.x. URL https://sfamjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-765X.2007.02237.x.

Mark Lindeman and Philip Stark. A gentle introduction to risk-limiting audits. *Security & Privacy, IEEE*, 10:42–49, 09 2012. doi: 10.1109/MSP.2012.56.

Sharon L Lohr. *Sampling: design and analysis*. CRC press, 2021.

Glen Meeden. A bayesian solution for a statistical auditing problem. *Journal of the American Statistical Association*, 98(463):735–740, 2003.

L.W. Miratrix and P.B. Stark. Election audits using a trinomial bound. *IEEE Transactions on Information Forensics and Security*, 4(4):974 – 981, 2009. URL http://ieeexplore.ieee.org/document/5282601/?reload=true.

G.P. Patil. *Composite sampling*, volume 1st Ed. A.H. El-Shaarawi and W.W. Piegorsch, 2002.

Philip B Stark. Conservative statistical post-election audits. *The Annals of Applied Statistics*, 2(2):550–581, 2008.

John P Wendell and Josef Schmee. Exact inference for proportions from a stratified finite population. *Journal of the American Statistical Association*, 91(434):825–830, 1996.

Appendix A

R Code

```
#preliminary initializing
#N = c(17000) # total batches
#contlevel = c(2.5,5,10,12.5) # levels of prevalent contamination
#alpha = .1
# N = Total number of batches (scalar or vector)
# contlevel = level of assumed contamination as a
# percent number (ie 1% contamination, contlevel = 1) (scalar or vector)
# alpha = sig level for hypothesis test

Batch = function(N,contlevel,alpha){
#storage for cont batches for each N
contbatforN = matrix(c(0),nrow = length(N), ncol = length(contlevel))
#storage for viable batches for each N
vbatch = matrix(c(0),nrow = length(N), ncol = length(contlevel))
for(i in 1:length(N)){
  contbatforN[i,] = N[i]*(contlevel/100) # contaminated batches
  vbatch[i,] = N[i] - contbatforN[i,]
```

```
   # vaible batches
}

# create master list to store p-values of each N for all contamination levels
master = vector(length = length(N),mode = 'list')
#store each probablilty matrix withing the associated Batch size
for(k in 1:length(N)){
  samplesize = seq(1:N[k])
  probmat = matrix(c(0),nrow = length(samplesize), ncol = length(contlevel))
   for (s in 1:length(samplesize)){
      probmat[s,] = phyper(0,contbatforN[k,],vbatch[k,],samplesize[s])
  master[[k]] = probmat
  }
}
master[[1]]

#################
# find the number of batches toe reach the confidence level #
#################
# for each element in master list,
#we want the first instance of <alpha for each cont level
#nmatrix is n batch storage for all total batch N
nmatrix = matrix(c(0),nrow=length(N), ncol = length(contlevel))
#nb is batch is row for a specifice Total batch N
#columns are contamination levels
nb = c(0)
for( j in 1:length(master)){
```

```
  for( i in 1:length(contlevel)){
    nb[i] = min(which(master[[j]][,i]<alpha))
  }
nmatrix[j,] = nb
}

colnames(nmatrix) = paste(contlevel,'% Contaminated', sep = '')
rownames(nmatrix) = paste(N,'Total Batches', sep = ' ')
nmatrix
}
# N = Total number of batches (scalar or vector)
# contlevel = level of assumed contamination (scalar or vector)
#alpha = sig level for hypothesis test.

contlevel = c(2.5,5,7.5,10,15,20)
N =  c(30,50,100)
##### checking ####
N = 500
contamlevel = c(2.5,5,10,20)
cbat = N*(contamlevel/100)
vbat = N - cbat
vbat

samplesize = seq(1,1000,1)
mat = matrix(c(0),nrow = length(samplesize),ncol = length(contamlevel))
for(i in 1: length(samplesize)){
  mat[i,] = phyper(0,cbat,floor(vbat),samplesize[i])
```

```
}
mat
n = c(0)
for(i in 1:ncol(mat)){
  n[i] = min(which(mat[,i]<.1))
}
n
#####################
# comparing with Jarvis #
#####################
# binomial => (1-p)^n when k = 0
# where p is assumed contamination
bin = seq(1,182,1)
ndraw = matrix(c(0), nrow = length(bin), ncol = length(contlevel))
p = .05
for(i in 1:length(bin)){
  ndraw[i,] = (1-(contlevel/100))^bin[i]
}
colnames(ndraw) = paste(contlevel,'% contamination')
rownames(ndraw) = bin

#90 conf
n90 = c(0)
for(i in 1:length(contlevel)){
  n90[i] = min(which(ndraw[,i]<.1))
}
```

```
#95 conf
n95 = c(0)
for(i in 1:length(contlevel)){
  n95[i] = min(which(ndraw[,i]<.05))
}

#99 conf
n99 = c(0)
for(i in 1:length(contlevel)){
  n99[i] = min(which(ndraw[,i]<.01))
}
binomsolution = rbind(n90,n95,n99)
colnames(binomsolution) = paste(contlevel,'% contamination')
rownames(binomsolution) = paste(c(90,95,99),'CL (BIN)')

##### PLOT COMPARING HYPER AND BINOM
###### DIFFERENT BATCH SIZES
##### NOTE BINOMIAL SOLUTION Is NOT AFFECTS BY N
#par(mfrow = c(1,1))
#N = 30
#90cl
H90 = Batch(N = 30,contlevel = contlevel, alpha = .1)
#95cl
H95 = Batch(N = 30,contlevel = contlevel, alpha = .05)
#99cl
H99 = Batch(N = 30,contlevel = contlevel, alpha = .01)
hypersol30 = rbind(H90,H95,H99)
```

```
rownames(hypersol30) = paste(c(90,95,99),'CL (HYP)')


#N = 30 batches graph
#hyper
plot(contlevel,hypersol30[1,], ylim = c(10,200),main = 'N = 30',
     xlab = 'Assumed Contamination', cex.lab = 1.5,
     ylab = 'Number of Samples Required',pch = 1,
     xaxt = 'n',yaxt = 'n',col = 'darkorchid3',lwd = 2)#90clhyp
lines(contlevel,hypersol30[1,], lty = 1, col ='darkorchid3',lwd = 2) # 90clhyp
points(contlevel,hypersol30[2,], pch = 2,col = 'darkorchid3',lwd = 2)#95clhyp
lines(contlevel,hypersol30[2,],lty = 2,col='darkorchid3',lwd=2)#95clhyp
points(contlevel,hypersol30[3,],pch = 3,col = 'darkorchid3',lwd = 2)#99clhyp
lines(contlevel,hypersol30[3,], lty = 3, col ='darkorchid3',lwd=2)#99lhyp
#Binomial
lines(contlevel,binomsolution[1,], lty=1, col = 'black',lwd=2)
points(contlevel,binomsolution[1,], pch=1, col = 'black',lwd = 2)
lines(contlevel,binomsolution[2,], lty=2, col = 'black',lwd=2)
points(contlevel,binomsolution[2,], pch=2, col = 'black',lwd = 2)
lines(contlevel,binomsolution[3,], lty=3, col = 'black',lwd=2)
points(contlevel,binomsolution[3,], pch=3, col = 'black',lwd = 2)
axis(1,at = contlevel, labels = paste0( contlevel,'%'),cex.axis = 1)
axis(2,at = seq(20,200,10),las = 1, cex.axis = 1)
legend(12,180, legend = c('90%','95%','99%'),
cex = 1.25,lty =c(1,2,3),lwd = 2 ,
       pch = c(1,2,3),title = 'Confidence Levels')
legend(12,120, legend = c('Binomial (Jarvis, 2007)',
('Hypergeometric')),cex = 1.25,
```

```
       fill =c('black','darkorchid3'),
       title = 'Method')

#N = 50
#90cl
H90 = Batch(N = 50,contlevel = contlevel, alpha = .1)
#95cl
H95 = Batch(N = 50,contlevel = contlevel, alpha = .05)
#99cl
H99 = Batch(N = 50,contlevel = contlevel, alpha = .01)
hypersol50 = rbind(H90,H95,H99)
rownames(hypersol50) = paste(c(90,95,99),'CL (HYP)')

#N = 50 batches graph
#hyper
plot(contlevel,hypersol50[1,], ylim = c(10,200),main = 'N = 50',
     xlab = 'Assumed Contamination',cex.lab = 1.5,
     ylab = 'Number of Samples Required',pch = 1,
     xaxt = 'n' ,yaxt = 'n',col = 'darkorchid3',lwd = 2)#90clhyp
lines(contlevel,hypersol50[1,], lty = 1, col ='darkorchid3',lwd = 2) # 90clhyp
points(contlevel,hypersol50[2,], pch = 2,col = 'darkorchid3',lwd = 2)#95clhyp
lines(contlevel,hypersol50[2,],lty = 2,col='darkorchid3',lwd = 2)#95clhyp
points(contlevel,hypersol50[3,],pch = 3,col = 'darkorchid3',lwd = 2)#99clhyp
lines(contlevel,hypersol50[3,], lty = 3, col ='darkorchid3',lwd = 2)#99lhyp
#Binomial
lines(contlevel,binomsolution[1,], lty=1, col = 'black',lwd = 2)
points(contlevel,binomsolution[1,], pch=1, col = 'black',lwd = 2)
```

```
lines(contlevel,binomsolution[2,], lty=2, col = 'black',lwd = 2)
points(contlevel,binomsolution[2,], pch=2, col = 'black',lwd = 2)
lines(contlevel,binomsolution[3,], lty=3, col = 'black',lwd = 2)
points(contlevel,binomsolution[3,], pch=3, col = 'black',lwd = 2)
axis(1,at = contlevel, labels = paste0( contlevel,'%'),cex.axis = 1)
axis(2,at = seq(20,200,10),las = 1, cex.axis = 1)
legend(12,180, legend = c('90%','95%','99%'),cex = 1.25,
lty =c(1,2,3),lwd = 2 ,
       pch = c(1,2,3),title = 'Confidence Levels')
legend(12,120, legend = c('Binomial (Jarvis, 2007)',
('Hypergeometric')),cex = 1.25,
       fill =c('black','darkorchid3'),
       title = 'Method')



#N = 100
#90cl
H90 = Batch(N = 100,contlevel = contlevel, alpha = .1)
#95cl
H95 = Batch(N = 100,contlevel = contlevel, alpha = .05)
#99cl
H99 = Batch(N = 100,contlevel = contlevel, alpha = .01)
hypersol100 = rbind(H90,H95,H99)
rownames(hypersol100) = paste(c(90,95,99),'CL (HYP)')

#N = 100 batches graph
```

```
#hyper
plot(contlevel,hypersol100[1,], ylim = c(10,200),main ='N = 100',
     xlab = 'Assumed Contamination',cex.lab = 1.5,
     ylab = 'Number of Samples Required',pch = 1,
     xaxt = 'n' ,yaxt = 'n',lwd = 2,col = 'darkorchid3')#90clhyp
lines(contlevel,hypersol100[1,], lty = 1, col ='darkorchid3',lwd = 2) # 90clhyp
points(contlevel,hypersol100[2,], pch = 2,col = 'darkorchid3',lwd = 2)#95clhyp
lines(contlevel,hypersol100[2,],lty = 2,col='darkorchid3',lwd = 2)#95clhyp
points(contlevel,hypersol100[3,],pch = 3,col = 'darkorchid3',lwd = 2)#99clhyp
lines(contlevel,hypersol100[3,], lty = 3, col ='darkorchid3',lwd = 2)#99lhyp
#Binomial
lines(contlevel,binomsolution[1,], lty=1, col = 'black',lwd = 2)
points(contlevel,binomsolution[1,], pch=1, col = 'black',lwd = 2)
lines(contlevel,binomsolution[2,], lty=2, col = 'black',lwd = 2)
points(contlevel,binomsolution[2,], pch=2, col = 'black',lwd = 2)
lines(contlevel,binomsolution[3,], lty=3, col = 'black',lwd = 2)
points(contlevel,binomsolution[3,], pch=3, col = 'black',lwd = 2)
axis(1,at = contlevel, labels = paste0( contlevel,'%'),cex.axis = 1)
axis(2,at = seq(20,200,10),las = 1, cex.axis = 1)
legend(12,180, legend = c('90%','95%','99%'),cex = 1.25,
lty =c(1,2,3),lwd = 2 ,
       pch = c(1,2,3),title = 'Confidence Levels')
legend(12,120, legend = c('Binomial (Jarvis, 2007)',
('Hypergeometric')),cex = 1.25,
       fill =c('black','darkorchid3'),
       title = 'Method')
```

```
####### 1500 = N


#90cl
H90 = Batch(N = 1500,contlevel = contlevel, alpha = .1)
#95cl
H95 = Batch(N = 1500,contlevel = contlevel, alpha = .05)
#99cl
H99 = Batch(N = 1500,contlevel = contlevel, alpha = .01)
hypersol1500 = rbind(H90,H95,H99)
rownames(hypersol1500) = paste(c(90,95,99),'CL (HYP)')


#N = 1500 batches graph
#hyper
plot(contlevel,hypersol1500[1,], ylim = c(10,200),main = 'N = 1500',
     xlab = 'Assumed Contamination',cex.lab = 1.5,
     ylab = 'Number of Samples Required',pch = 1,xaxt = 'n' ,
     yaxt = 'n',lwd = 2,col = 'darkorchid3')#90clhyp
lines(contlevel,hypersol1500[1,], lty = 1, col ='darkorchid3',
lwd = 2) # 90clhyp
points(contlevel,hypersol1500[2,], pch = 2,col = 'darkorchid3',
lwd = 2)#95clhyp
lines(contlevel,hypersol1500[2,],lty = 2,col='darkorchid3',
lwd = 2)#95clhyp
points(contlevel,hypersol1500[3,],pch = 3,col = 'darkorchid3',
lwd = 2)#99clhyp
lines(contlevel,hypersol1500[3,], lty = 3, col ='darkorchid3',
lwd = 2)#99lhyp
```

```
#Binomial
lines(contlevel,binomsolution[1,], lty=1, col = 'black',lwd = 2)
points(contlevel,binomsolution[1,], pch=1, col = 'black',lwd = 2)
lines(contlevel,binomsolution[2,], lty=2, col = 'black',lwd = 2)
points(contlevel,binomsolution[2,], pch=2, col = 'black',lwd = 2)
lines(contlevel,binomsolution[3,], lty=3, col = 'black',lwd = 2)
points(contlevel,binomsolution[3,], pch=3, col = 'black',lwd = 2)
axis(1,at = contlevel, labels = paste0( contlevel,'%'),cex.axis = 1)
axis(2,at = seq(20,200,10),las = 1, cex.axis = 1)
legend(12,180, legend = c('90%','95%','99%'),cex = 1.25,
lty =c(1,2,3),lwd = 2 ,
       pch = c(1,2,3),title = 'Confidence Levels')
legend(12,120, legend = c('Binomial (Jarvis, 2007)',
('Hypergeometric')),cex = 1.25,
       fill =c('black','darkorchid3'),
       title = 'Method')

#### tables for the report option
## for excel



rownames(hypersol30) = paste(c(90,95,99),'% Confidence',sep='')
rownames(hypersol50) = paste(c(90,95,99),'% Confidence',sep='')
rownames(hypersol100) = paste(c(90,95,99),'% Confidence',sep='')
rownames(hypersol1500) = paste(c(90,95,99),'% Confidence',sep='')

hypersol30 = as.data.frame(hypersol30,
```

```
cbind(c(paste(c(90,95,99),'% Confidence',sep='')))
hypersol50= as.data.frame(hypersol50)
hypersol100= as.data.frame(hypersol100)
hypersol1500= as.data.frame(hypersol1500)
```