Changing the Set Sizes in Raw Ground Poultry Sampling

United States Department of Agriculture

Food Safety and Inspection Service

August 2012

Executive Summary

This report provides a summary of a statistical analysis conducted to examine the sampling set sizes for raw ground poultry.[1,2] Section 1 examines the current state of raw ground poultry sampling. Section 2 presents the proposed changes to raw ground poultry sampling programs.[3] The details of the statistical analysis are contained in the appendices.

This analysis was conducted by reviewing FSIS' documentation and sampling data for raw ground poultry from fiscal years 2009-2011 (FY09-11). During the three year period, samples were collected for 74 sets, which were conducted in 38 establishments. Using PHIS data, we now know that 140 establishments are eligible for raw ground poultry sampling sets under current criteria. Of these 74 sets, 51 contained at least 50 samples.[4]

The current FSIS performance standards were set by using the results of baseline sampling data, and were set so establishments performing at the performance standard have an 80% probability of passing. This analysis reviewed the sampling results in relation to the performance standards, and found that only one set had a *Salmonella* positive rate that was higher than the performance standard.

Analysis was performed on the collection properties of samples within a set. The average length for sets was 183 and 154 days in chicken and turkey respectively. The analysis revealed that positive samples were randomly distributed across sets, which means that the sampling data are independent across time and that decreasing the sampling window should produce the same results when comparing an establishment's sampling results to performance standards.

In its Strategic Data Analysis Plan (October 2010)[5], FSIS committed to looking for alternative approaches to sampling design. This report presents analysis and recommendations to modify the raw ground poultry sampling sets.

The effect of reducing the set size was explored by using historical data and FSIS' existing methodology to generate the maximum number of allowable positives for different set sizes. Historical sampling sets were also reviewed to determine if and how the outcomes would change if the set size was lowered. Analyzing the upper confidence bound shows that set sizes as small as 20 were adequate to assess whether a set passed or not. Confidence intervals were also used to investigate a change from 30 to 50 samples. This analysis found that increasing the set size from 30 to 50 samples usually did not cause the set that passed at 30 samples to exceed the failing threshold at 50.

---

[1] A set is a collection of samples collected on consecutive days of production at a single establishment.

[2] This report draws heavily from the report "*Salmonella* Sampling Set Sizes in Raw Ground Poultry Products, FY09-11" that was developed by the MITRE Corporation in March 2012 under contract with the Data Analysis and Integration Group (DAIG) within the Office of Data Integration and Food Protection (ODIFP). That report contained a majority of the analyses and is the source of the recommendations.[3] MITRE made the recommendations in Section 2 and FSIS is considering their recommendations.

[3] MITRE made the recommendations in Section 2 and FSIS is considering their recommendations.

[4] It was determined that 50 samples were sufficient to include a set in this analysis. Only one set had a *Salmonella* positive rate that was above the maximum "pass" level.

[5] This report can be found on the FSIS website at:
http://www.fsis.usda.gov/OPPDE/NACMPI/Sep2010/2010_Strategic_Data_Analysis_Plan.pdf .

This paper recommends consideration of a change in sample set sizes to free up sampling resources. There are multiple benefits to implementing this recommendation. First of all, this change, as recommended, would increase the number of available sets, which in turn would increase the number of establishments being sampled under FSIS' *Salmonella* testing program at any given time. Also, the increase in testing may influence establishments to adopt procedures to improve food safety. The drawback of this change is that it would slightly decrease the statistical power of the sampling test. This means that there would be a small increase in the chance that a noncompliant establishment, one which is operating worse than the performance standard, would pass the set.

With the increased frequency of sampling, FSIS would be better able to identify noncompliant establishments. Furthermore, the likelihood of a noncompliant establishment not being identified as performing worse than the standard is 100% if the establishment is not sampled at all. In addition, the ability to correctly detect compliant establishments would be at about the same level as it is currently. It is important to note that this recommendation is based on real-world considerations and while it is statistically informed, it is not a statistical recommendation.

Acronyms

03B	HACCP code for “Raw Ground Poultry”
03C	HACCP code for “Raw Not Ground Poultry”
03J	HACCP code for “Poultry Slaughter”
DAIG	Data Analysis and Integration Group
FSIS	Food Safety and Inspection Service
FSA	Food Safety Assessment
HACCP	Hazard Analysis and Critical Control Points
NR	Non-compliance record
ODIFP	Office of Data Integration and Food Protection
PBIS	Performance Based Inspection System
PHIS	Public Health Information System
StDev	Standard Deviation
USDA	United States Department of Agriculture
W3NR	Public Health related non-compliance record

Table of Contents

List of Figures

List of Tables

1. *Salmonella* Sampling of Poultry at FSIS

The Food Safety and Inspection Service (FSIS) within the United States Department of Agriculture (USDA) is responsible for ensuring that the nation's commercial supply of meat, poultry, and processed egg products is safe, wholesome and correctly labeled and packaged. This report first reviews historical and current FSIS *Salmonella* sampling procedures and activities, then provides major findings from the analysis regarding the Agency's *Salmonella* sampling set size and finally provides the recommendations regarding this analysis.

1.1 Introduction

The overall purpose of FSIS inspection and sampling is to ensure that establishments maintain control of their production processes and adhere to FSIS regulations, policies and performance standards, which the Agency believes helps protect the public from foodborne illnesses. Product testing, whether performed by industry or FSIS, is particularly important in gauging the safety of regulated product. The routine sampling in FSIS-regulated domestic and import establishments allows the Agency to assess the effectiveness of industry process controls, compliance with performance standards and the monitoring the proportion of finished product where microbiological or chemical contaminants are detected on products being produced for American consumers. Additionally, sampling serves as a strong incentive for the meat, poultry and processed egg product industries to reduce the presence of pathogens on products they produce. Further, product sampling provides the regulated industries with critical information to improve current processes and focus their resources as efficiently and effectively as possible.[6]

More specifically, FSIS collects samples of products from establishments throughout the year to undergo Pathogen Verification Testing. Different sampling programs focus on different pathogens and products.

The raw poultry sampling program currently covers the following types of poultry: post-chill young chicken, post-chill young turkeys, raw ground chicken and raw ground turkey.[7] The raw ground poultry sampling program currently focuses on *Salmonella*.

Not all poultry establishments are sampled, and those that undergo sampling are sampled at different frequencies. The type of product produced and production volumes affect whether and how often FSIS collects samples from an establishment. In addition, sampling may be deferred for establishments with a history of few *Salmonella* positives and establishments may be excluded from sampling if they have a low production volume.[8]

[6] For more information about FSIS' sampling programs, please see the Agency's " Report on the Food Safety and Inspection Service's Microbiological and Residue Sampling Programs" at: http://www.fsis.usda.gov/PDF/FSIS_Sampling_Programs_Report.pdf

[7] At the time this report was written, the Salmonella sampling program did not cover chicken parts (intact pieces of chicken such as thighs or legs), mixed species raw ground poultry, mechanically separated raw poultry or marinated raw poultry.

[8] See FSIS' website for more information on the scheduling criteria: http://www.fsis.usda.gov/Science/Scheduling_Criteria_*Salmonella*_Sets/index.asp

1.2 Current Performance Standards

In 2010, FSIS published new *Salmonella* and *Campylobacter* performance standards for young chickens and young turkeys.[9] The technical report[10] written as part of that policy change (hereafter referred to as "the 2010 technical report") states that per the Pathogen Reduction/Hazard Analysis and Critical Control Points (PR/HACCP) Final Rule (1996)[11], FSIS' general methodology for determining performance standards for a pathogen in a product type consists of:

1) Carrying out a nationwide baseline survey (at least a year in length) to identify the prevalence of the pathogen in a specific product.
2) Selecting a sample set size that is greater than or equal to 50.
3) Selecting the maximum number of positives such that an establishment performing at the performance standard has roughly 80% probability of passing; or conversely, there is only a 20% chance of an establishment performing at the standard failing the test.

More concretely, the 1995 FSIS "Nationwide Raw Ground Chicken Microbiological Survey" estimated that the nationwide prevalence of *Salmonella* in raw ground chicken was 44.6%.[12] FSIS selected a sample size and then considered the false positive rate for a hypothetical establishment performing at the performance standard. For such a hypothetical establishment, FSIS wanted a 20% chance of a false positive. That is, for an establishment whose actual prevalence was at the performance standard, there was a 20% chance that it would have greater than the maximum number of positive samples in a given set.

The sample size for ground chicken and turkey has historically been 53. FSIS used the binomial distribution function to determine that if an establishment was performing at the performance standard, 78.6% (roughly 80%) of the probability distribution was below 49.1% (or 26 positives out 53). Therefore, FSIS set 26 as the maximum number of *Salmonella* positives in a 53 sample set to count as passing for raw ground chicken. In Figure 1, the binomial probability distribution function is presented for an establishment with an actual prevalence of *Salmonella* at the

---

Also see "Standard Operating Procedures for *Salmonella* and *Campylobacter* Verification Testing" 2011: http://www.fsis.usda.gov/PDF/SOP_*Salmonella*_Eligibility_Testing_092211.pdf.

[9] "New Performance Standards for Salmonella and Campylobacter in Young Chicken and Turkey Slaughter Establishments; New Compliance Guides", May 14, 2010. https://www.federalregister.gov/articles/2010/05/14/2010-11545/new-performance-standards-for-salmonella-and-campylobacter-in-young-chicken-and-turkey-slaughter

[10] "Draft: Technical paper for performance guidance for broilers and young turkey at post-chill." http://www.fsis.usda.gov/PDF/Technical_Paper_Performance_Guidance_Broilers.pdf

[11] See FSIS' website for more information at: http://www.fsis.usda.gov/OPPDE/rdad/FRPubs/93-016F.pdf.

[12] "Nationwide Raw Ground Chicken Microbiological Survey." 1995. Available: http://www.fsis.usda.gov/OPHS/baseline/rwgrchck.pdf. The "Nationwide Raw Ground Turkey Microbiological Survey" 1995 is available here: http://www.fsis.usda.gov/OPHS/baseline/rwgrturk.pdf. The chicken survey was carried out in only 30 establishments, and the turkey survey was carried out in only 40 establishments. Because of budgetary constraints, each survey was limited to a maximum of 300 samples.

performance standard. An establishment performing at the performance standard has roughly an 80% chance of receiving 26 or fewer positives in a sample set of 53 (the probability mass to the left of the red "Maximum Number of Positives" line).

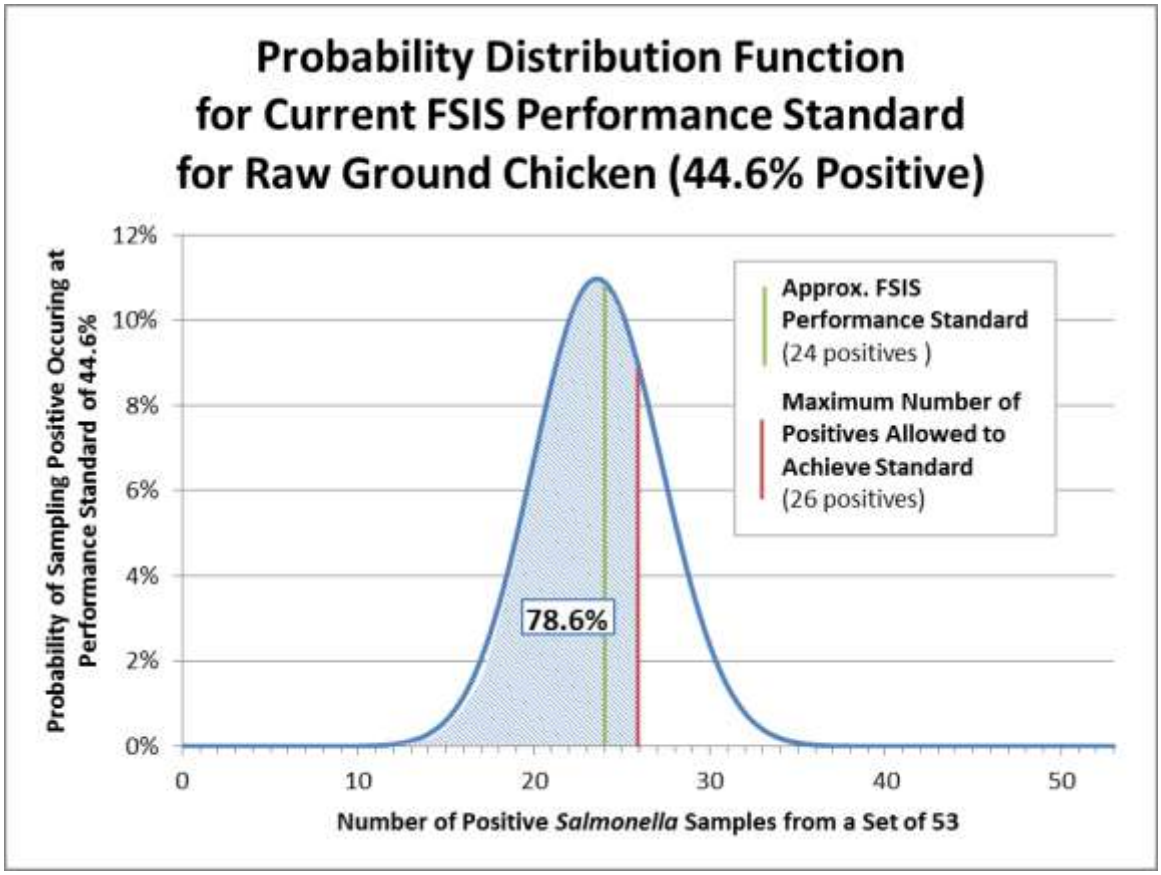

Figure 1: Binomial Distribution at the Current Performance Standard (44.6% in 53 Samples) for Raw Ground Chicken

In Table 1, the current performance standards for all four categories of poultry products tested in the *Salmonella* sampling program are presented.

Category	Performance Standard	Number of Samples in a Set	Maximum Number of Positives to Achieve Standard	Maximum Number Divided by Set Size	Probability of Passing the Criterion when an Establishment's *Salmonella* Positive Rate is at the Performance Standard
Young Chicken	7.5%	51	5	9.8%	81.9%
Turkey	1.7%	56	4	7.1%	99.7%
Ground Chicken	44.6%	53	26	49.1%	78.6%
Ground Turkey	49.9%	53	29	54.7%	79.9%

Table 1: *Salmonella* Performance Standards in Poultry Products[13]

It is important to note that turkey has a dramatically different probability of passing than other poultry products. The FSIS 2010 technical paper observes:

> As a result of the relatively low estimate of incidence, the Agency decided to adopt a sampling plan such that an establishment meeting the performance standard would have more than a 99.0% probability of passing the compliance criterion, rather than the 80% rule used for deriving a compliance criterion. The number of samples in a set is 56, thus, the compliance rule no more than 4 positive results within 56 samples satisfies this criterion with the least number of permitted positive results. For a performance standard of 1.7%, the probability of less than or equal 3 positive results is 98.48% which is less than 99.0%; the probability of less than or equal to 4 positive results is about 99.74%.[14]

1.3 Assessment of Raw Ground Poultry *Salmonella* Sets

This analysis was conducted by reviewing FSIS' documentation regarding raw ground poultry sampling and Agency *Salmonella* sampling data from fiscal year 2009-2011 (FY09-11). It is important to note that these analyses focused only on *Salmonella* sampling of raw ground poultry products. These analyses were primarily focused in three areas: Type I/Type II error, the number of eligible establishments, and the number of *Salmonella* sets conducted by FSIS during the study period.

[13] The first three columns of numbers derive from FSIS' "Q3 2011 Salmonella Testing Report", available: http://www.fsis.usda.gov/Science/Q3_2011_Salmonella_Testing_Tables/index.asp#table1a. The last two columns were computed by MITRE. The probabilities for "Broilers" and "Turkey" in the last column match those in "Draft: Technical paper for performance guidance for broilers and young turkey at post-chill."

[14] "Draft: Technical paper for performance guidance for broilers and young turkey at post-chill." Pages 55-56.

1.4 Number of Eligible Establishments

The number of establishments producing raw ground chicken or turkey products that are eligible for FSIS *Salmonella* verification testing can be estimated using the Agency's new Public Health Information System (PHIS).[15] FSIS uses a number of criteria to determine whether an establishment is eligible for testing.

First, an establishment is must produce a comminuted raw ground poultry product.[16] According to the establishment product table in PHIS as of August 2012, there were 700 establishments with non-null daily production volume estimates for "raw ground, comminuted, or otherwise non-intact" chicken or turkey, 583 for chicken, 258 for turkey.[17] Second, an establishment needs to produce a minimum amount of product on a daily basis, which decreases the number of eligible establishments from 700 to 247. Finally, at the time of this analysis, FSIS does not sample "mechanically separated" and "otherwise non-intact" product. This means that the number of eligible establishments decreases from 247 to 140.[18]

1.5 Eligible Establishments with *Salmonella* Sets During the Study Period

According to the FY09-11 sampling data, there were 74 sample sets for raw ground poultry products.[19] Only 51 of these sets had 50 or more samples. Of the 74, 28 were for raw ground chicken and the other 46 were for raw ground turkey. Over the three year period, only 38 establishments received *Salmonella* testing for raw ground poultry products.

If considering only those sample sets with 50 samples or more in FY09-11, only one set had a *Salmonella* positive rate that was above the maximum "pass" level. The current performance standard (based on 1995 data) is far higher than the percent positive in establishments that were tested in FY09-11.

At the FY09-11 sampling rate (roughly 25 sets per year), it would take slightly less than six years to sample all 140 establishments that are currently eligible just once. If the rate is calculated based on the number of sets with more than 50 samples (51 sets over three years=17 sets per year), it would take more than eight years.

1.6 Set Length

The length of time needed to complete a sampling set contributes to its effectiveness for assessing industry process controls. Part of the decision to make sets 50 or more samples was to ensure that sets lasted long enough to evaluate an establishment's process controls. When the set program was started, the assumption was that two months of data should be enough to determine

[15] Prior to PHIS, FSIS used static product information, which resulted in a smaller sampling frame.

[16] At the time of this analysis, "mechanically separated" and "otherwise non-intact" product were not sampled by FSIS. The current Federal Register Notice is removing this restriction.

[17] These numbers do not sum to 700 because some establishments process both chicken and turkey.

[18] After the current Federal Register Notice is finalized, the number of eligible establishments will be 247.

[19] As of August 2012, 54 ground poultry sets have been started in PHIS since it was started in April 2011.

if an establishment meets the performance standards or not. The analysis of Agency ground poultry *Salmonella* sampling data from fiscal year 2009-2011 (FY09-11) found that the average length for a ground chicken set was 183.1 days and the average length for a ground turkey set was 154.2 days. The median length for a chicken set was 116 days and the median length for a turkey set was 93 days.

1.7 Sampling Windows

In order to accurately assess the importance of sample size, it is necessary to confirm that the sampling data are independent over time (i.e., that samples taken at the end of a set are no more or less likely to be positive for *Salmonella* than samples taken at the beginning). This is exceptionally important given the long duration of most sets.

The raw ground poultry *Salmonella* sampling data from fiscal year 2009-2011 (FY09-11) were inspected both visually and statistically to verify that the samples were normally distributed.[20] Visual inspection of the data showed no anomalies or obvious trends. Positive samples appeared to be randomly distributed across all windows. Sets that had fewer than 50 samples for this portion of the analysis were ignored, as well as any samples after the 50th.

Additionally, for both chicken and turkey, a single-factor analysis of variance (ANOVA) was performed. The test checks to see if the average sampling rate in sample window 1-10 is the same as sample window 11-20, and so on. For both raw ground chicken and raw ground turkey, the analysis indicated that the sampling windows have little impact on the positive sampling rate. This implies that a shorter window should produce the same results when comparing an establishment's sampling results to the performance standard.

1.8 Change as Set Size Increases

To determine if changes occur as set size increases, the sampling data was viewed in a sequential, cumulative manner. For each set, the first five samples were used to generate a positive *Salmonella* rate. The next five samples were then used and a cumulative rate over the first ten samples was generated. This was then repeated for the first 50 samples of a set. Again, sets that had fewer than 50 samples were ignored. As before, there were 17 sets of raw ground chicken and 34 sets of raw ground turkey. Visual inspection suggests that the positive sampling rate begins to level off after 30 samples for both types of product. Out of 17 ground chicken sets, only 2 of them had a change in percent positive by over 10% after 30 samples. The same is true of the 34 ground turkey sets analyzed – only 2 sets had more than 10% change in positive rates after the 30th sample was collected.

[20] A normal distribution is important because that ensures that the samples are statistically independent. If the samples are statistically independent, then FSIS will be able to assess an establishment's process control with a smaller number of samples.

1.9 Balancing Type I/Type II Error

Type I error occurs when an establishment is recorded by FSIS as failing a sampling set when, in fact, they are operating within the given performance standards. Type I error is also known as the "false positive" rate. Conversely, Type II error occurs when an establishment passes a sampling set when, in fact, they are operating above the performance standard Type II error is also known as the "false negative" rate. It is important to recognize the Type I and Type II error rates in sampling programs as unavoidable realities of statistical design, which can only be minimized to a certain extent without compromising the time and money allocated to a project. That is, because of the inverse relationship between the two error types, with all other factors held constant (i.e. only changing the number of allowable max positives), one cannot decrease one error type without increasing the other. The analysis reviewed FSIS' performance standards for poultry to determine the impact of changing the sampling set sizes from the current size on Type I and Type II error.

Table 2 contains results from exact binomial distribution functions for different scenarios.

Set Size	Max Positives	Chance of False Positive -actual prevalence is 44.6%	Chance of False Negative -actual prevalence is 50%	Chance of False Negative -actual prevalence is 55%	Chance of False Negative -actual prevalence is 60%	Chance of False Negative -actual prevalence is 65%
5	3	12.7%	81.3%	74.4%	66.3%	57.2%
10	5	25.3%	62.3%	49.6%	36.7%	24.9%
15	8	17.3%	69.6%	54.8%	39.0%	24.5%
20	10	23.8%	58.8%	40.9%	24.5%	12.2%
25	13	17.2%	65.5%	45.7%	26.8%	12.5%
30	15	21.8%	57.2%	35.5%	17.5%	6.5%
35	18	16.3%	63.2%	39.8%	19.3%	6.8%
40	20	19.8%	56.3%	31.6%	13.0%	3.6%
45	22	23.2%	50.0%	24.9%	8.6%	1.9%
50	25	18.1%	55.6%	28.4%	9.8%	2.1%
53	26	21.4%	50.0%	23.2%	7.0%	1.2%
55	27	21.0%	50.0%	22.8%	6.6%	1.1%
60	30	16.6%	55.1%	25.8%	7.5%	1.2%
65	32	19.0%	50.0%	20.9%	5.1%	0.6%
70	34	21.5%	45.2%	16.8%	3.5%	0.3%
75	37	17.3%	50.0%	19.2%	4.0%	0.4%

Table 2: Example of Type I and Type II Errors for Different Set Sizes

Each row of Table 2 represents different *Salmonella* set sizes, ranging from 5 to 75; a row has also been included for a 53 sample set size, the current set size for raw ground chicken and turkey. For each set size, FSIS' desired false positive rate was approximated (because the binomial distribution is discrete, some of numbers are farther from 20% than would be ideal) for a performance standard of 44.6%.

In the green and red columns, probabilities were recorded for establishments whose actual (but unknown) prevalence is the number in parentheses.[21] For example, in the green column, probabilities were recorded for an establishment whose actual prevalence is 44.6%. In the first red column, probabilities were recorded for an establishment whose actual prevalence is 50%. The difference between the green and the red columns is that in the green column, the estimated false positive rate (set declared a "fail" when an establishment is actually at the performance standard) is recorded. In the red column, the estimated false negative rate (set declared a "pass" when an establishment's actual prevalence is above the performance standard) is recorded. So, if an establishment has an actual (but unknown) prevalence of 44.6% and the set size is 53, there is a 21.4% chance that it will incorrectly fail a set. If the actual prevalence is 55%, there is a 23.2% chance that it will incorrectly pass the test (or a 76.8% chance it will be detected).

To illustrate the effect of reducing the set size, the false negative rates for set sizes of 30 samples should be compared to the false negative rates for set sizes of 53 samples. The probability of FSIS incorrectly passing a non-compliant establishment with an actual *Salmonella* prevalence of 44.6% would increase from 50% to 57.2%.

According to the FSIS FY09-11 *Salmonella* sampling data, only one set with 50 or more samples failed and it only exceeded the "maximum number divided by the set size" (see Table 2) by one positive sample (28 positives in a 55 sample set yields an estimated prevalence of 50.9%). Using the binomial distribution function, the chance of this set passing the chicken standard (having 26 or fewer positives) with a set size of 53 is 45%, and the chance of it passing the chicken standard (having 15 or fewer positives) with a set size of 30 is 53%.

2. Recommendations

Based on the analyses conducted, a list of recommendations was developed.[22] This list is provided below.

1) FSIS should consider determining its set size and performance thresholds based on desired operating characteristics (desired false positive and false negative rates) over a range of different hypothetical rates of prevalence at a given establishment.
2) If FSIS' main concern is to ensure that an establishment at the performance standard has an 80% chance of passing a *Salmonella* set, then FSIS should consider lowering the set size to 30 samples. This will increase the number of available sets by 46%. FSIS will need to balance this decision with the effect that a decrease in sample size would increase the rate of false negatives (establishments whose actual prevalence is above the performance standard who "pass" a set). However, those establishments not being sampled or have never been sampled are guaranteed not to fail.
3) To achieve FSIS' goal of sampling every raw ground poultry establishment twice in a timely manner, major changes to the current program need to be made. These might include:

[21] The actual prevalence cannot be known with 100% certainty as FSIS does not test every single bird produced by all establishments. Therefore, the actual, true prevalence is unknown.

[22] MITRE made these recommendations in the report *Salmonella* Sampling Set Sizes in Raw Ground Poultry Products, FY09-11" that was developed by the MITRE Corporation in March 2012. FSIS is considering these recommendations.

a. Shifting the allotment of sets from products with lower rates of *Salmonella* to raw ground poultry.
b. Shortening the sample set size.
c. Obtaining more resources to increase the number of *Salmonella* sampling tests.

4) If FSIS is unable to sample all eligible raw ground poultry establishments, it should consider adding a random sampling program to extend the Agency's influence with the current resources available.
5) FSIS should consider aligning the performance standard for raw ground poultry products more closely to current nationwide prevalence.
6) An alternative to simply reducing the set size would be to use variable set sizes. Stopping conditions were generated at 20, 30, and 40 samples. Ten samples were ignored because the confidence interval was exceptionally wider than at 20 samples. **Error! Reference source not found.** shows the suggested stopping points at 20, 30, 40, and 50 samples. Since turkey sets had lower positive rates in this analysis than chicken sets on average, bounds for all poultry were established based on the current FSIS failing threshold for chicken. The "Stop Below" column indicates the highest observed percentage at which the 95% upper bound is below the failing threshold for raw ground chicken sets. If a set had 25% (5/20) positives after 20 samples, the set would end with an automatic pass. If it had 30% (6/20), it would continue, and so on. The "Stop Above" column is the lowest observed percentage at which the 95% lower confidence bound remains above the failing threshold. Any observed percentage above the stopping condition would result in an automatic fail, though this condition is less likely to be needed; only one of the 51 sets reviewed would have triggered a stopping condition.

Sample Size	Stop Below	Stop Above
20	**28%**	**70%**
30	**32%**	**66%**
40	**34%**	**64%**
50	**36%**	**62%**

Table 3: Suggested Set Stopping Conditions Based on One-Sided 95% Confidence Bounds.

If variable set sizes are used, 50 samples should be used at the maximum set size, and 36% should be used as the failing threshold. This ensures a 95% confidence that any passing set is below the current 49.1% threshold. Beyond this, the number of samples required to continue increasing precision grows exponentially and therefore would not be an effective use of resources. A slightly more complex method of shortening a set size based on observed performance would be the Sequential Probability Ratio Test (SPRT).

3. Appendices

3.1 Introduction

These appendices contain additional analyses conducted on the FY09-11 raw ground poultry *Salmonella* sampling data to support the conclusions and recommendations above.

In many cases, this report displays findings in box-plots, a common tool in statistics to visually display results. Typically, box plots splits the data set into quartiles. The body of the box plot consists of a "box", which goes from the first quartile (the top of the box) to the third quartile (the bottom of the box). Within the box, a vertical line is drawn inside the box, which shows the median of the data set. In this situation, the decision was made to display both the "average" and the "median". Here, the average is shown as a vertical dotted line, whereas the median is shown as a vertical solid line. Additionally, two horizontal lines, called whiskers, extend from the front and back of the box. The front whisker goes from top of the box to the smallest non-outlier in the data set, and the back whisker goes from the bottom of the box to the largest non-outlier. If the data set includes one or more outliers, they are plotted separately as points on the chart. Each box plot also contains information about the number of establishments that were included in each analysis. This number is represented by the letter "n" and is typically displayed at the top of each "box".

3.2 Number of Samples per Set

Most of the sets analyzed did not contain the FSIS standard of 53 samples.[23] Figure 2 shows a box plot of the set size for both chicken and turkey, but it does not account for the possibilities of outliers. The plots show the minimum, lower quartile, median, upper quartile, maximum, and the number of sets included. The plot also shows the average number of samples—44.6 for chicken sets and 47.4 for turkey sets—as the dotted lines. The median number of samples for chicken sets was 53 and the median number of samples for turkey sets was 54.5. Both chicken and turkey have long lower whiskers and the average for turkey is actually below the lower quartile. This is due to three very small sets (with 5, 11, and 12 samples).

[23] Prior to PHIS, FSIS used the PREP system to schedule samples. The PREP system relied on paper forms that were mailed to FSIS inspectors in the field. Because of latencies in that system, FSIS scheduled extra samples for each set to ensure that enough samples were collected to complete the set. Only the first 53 valid samples were used by FSIS to determine if an establishment passed or failed a set. In PHIS, the exact number of sampling tasks needed to complete a set is assigned. Then, additional samples are added if necessary to replace discarded samples.

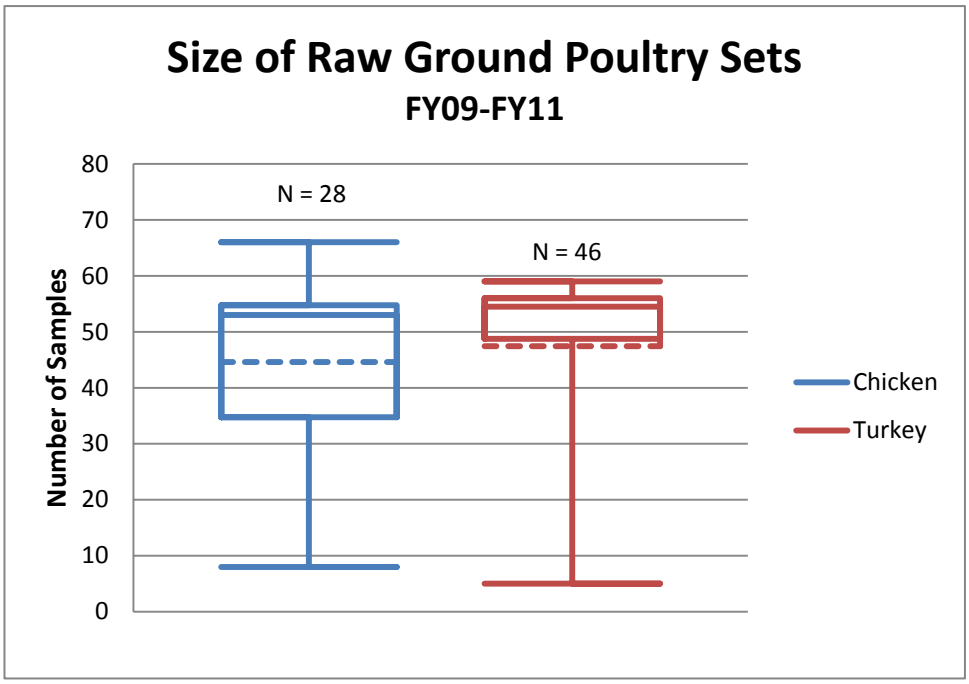

Figure 2: Size of Raw Ground Poultry Sets (Number of Samples Taken) for FY09-FY11.

3.3 Length of Sample Sets in Days

Figure **3** shows a box plot for the length in days of both raw ground chicken and raw ground turkey sets. The average length for chicken was 183.1 days and the average for turkey was 154.2 days. The median length for chicken was 116 days and the median length for turkey was 93. Here, the upper whiskers are longer, skewing the mean above the median. However, rather than treat exceptionally long sets as outliers, the effect of time on sampling sets was investigated—this is discussed below. By combining the length and size of the sampling sets, the average pace of raw ground chicken sets was determined to be 4.16 days per sample and 3.79 days per sample for turkey. If the median was used, the pace for raw ground chicken was 2.65 days per sample and 1.77 days per sample for raw ground turkey.

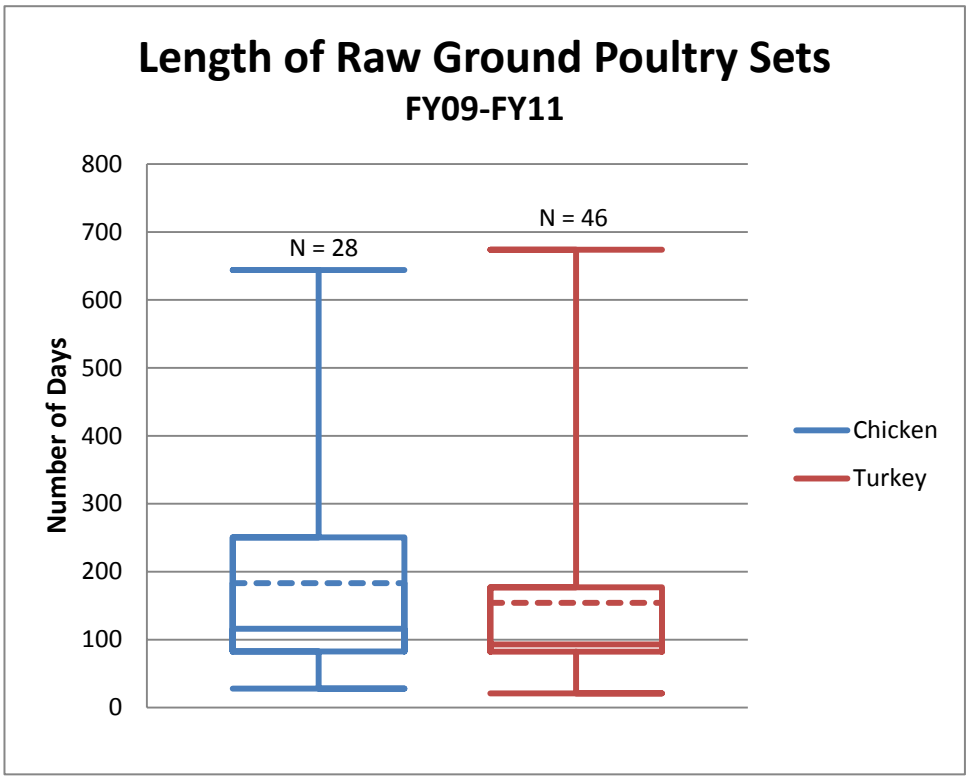

Figure 3: Length of Raw Ground Poultry Sets in Days for FY09-FY11.

3.4 Differences between First and Second Sets per Establishment in FY09 to FY11

Analyses were conducted to determine how performance changed between multiple sets at the same establishment. There were 38 unique establishments sampled from FY09 to FY11. Of these, 16 were sampled for raw ground chicken and 23 were sampled for raw ground turkey (one establishment was sampled for both chicken and turkey). Three chicken establishments were sampled in three sets and another six establishments had two sets. The rest were only sampled once. Of the nine establishments with multiple sets, five establishments had a lower positive sampling rate in their last set than in their first set. The average change for raw ground chicken establishments was -1.1% between the first and last set. The outcomes were similar for turkey; of 23 establishments, 19 had multiple sets. Ten establishments had a lower positive sampling rate in the last set than in the first set while eight had a higher rate (one establishment sampled at 0% in both sets so had no change between the two). The average difference was -1.4%. With so few observations it is difficult to draw conclusions, but this finding suggests that on average, establishments do not perform either better or worse in a follow-on set.

3.5 Confidence Intervals on Sampling Sets

Confidence intervals are dependent upon sample size. A larger sample size allows for a narrower confidence interval (i.e., more precision). The confidence interval for a binomial distribution is also dependent upon the observed percentage (in this case, the positive sampling

rate). A 95% confidence interval for an establishment with a 10% positive sampling rate will be narrower than one for an establishment with a 30% positive rate. Binomial confidence intervals are widest when the observed percentage is 50%. Figure 4 shows the two-sided confidence intervals (up to an observed percentage of 50%) for samples sizes of 10, 20, 30, 40 and 50 samples.

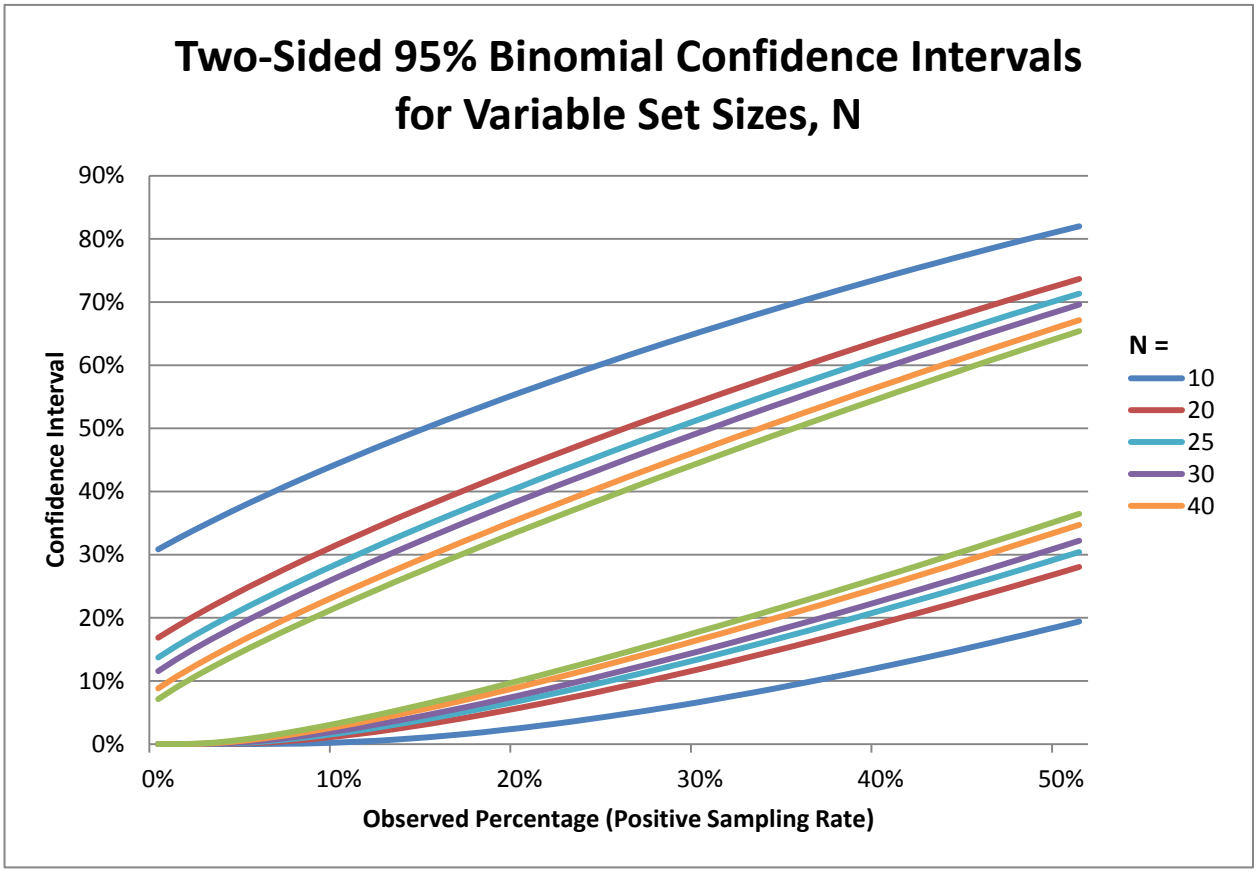

Figure 4: Two-Sided 95% Binomial Confidence Intervals for Variable Set Sizes (N=10, 20, 30, 40, and 50).

If the sample set size is 10 and the observed percentage is 10%, there is a 95% chance that the actual percentage falls somewhere between 1% and 44%. If set size increases to 20, and the observed percentage is 10%, there is a 95% chance that the actual percentage falls between 2% and 31%. Finally if the set size is increased to 50 and the observed percentage is 10%, there is a 95% chance that the actual percentage falls between 3% and 21%.

As the sample set sizes increase, the confidence interval becomes more precise and the width decreases, but this difference becomes smaller as the sample size continues to increase. Figure 5 demonstrates this diminishing return with 95% confidence intervals. At low set sizes, all the widths are large, but drop off rapidly as the set size is increased. In fact, the width has an asymptotic relationship with zero. This means the width of a binomial confidence interval will approach, but never reach, zero as the set size moves out to infinity.

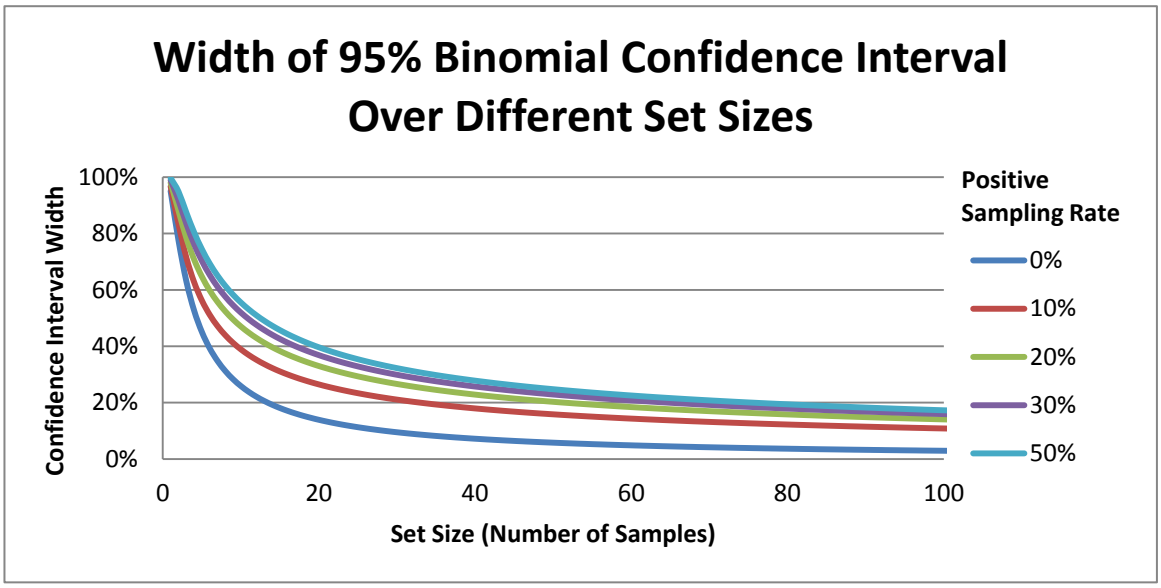

Figure 5: 95% Upper Confidence Bound Over Variable Sample Sizes.

To gain a sense of the distributions of the percent positives in *Salmonella* sampling sets, each set with 95% confidence intervals was plotted as a point. Exact binomial confidence interval calculations were used for this data, and confidence intervals were calculated for all sampling sets in the data. Figure 6 shows the 28 chicken sets and Figure 7 shows the 46 turkey sets. As explained at the beginning of this section, the differently-sized confidence intervals were due to the observed percentage (positive sampling rate) and the variable set sizes. The larger confidence intervals were driven mostly by smaller set sizes. The ordering of sets on the x-axis was based on the value of the y-axis; it was not temporal. The performance standard is shown as a green line and the maximum allowable percentage is shown as a red line.

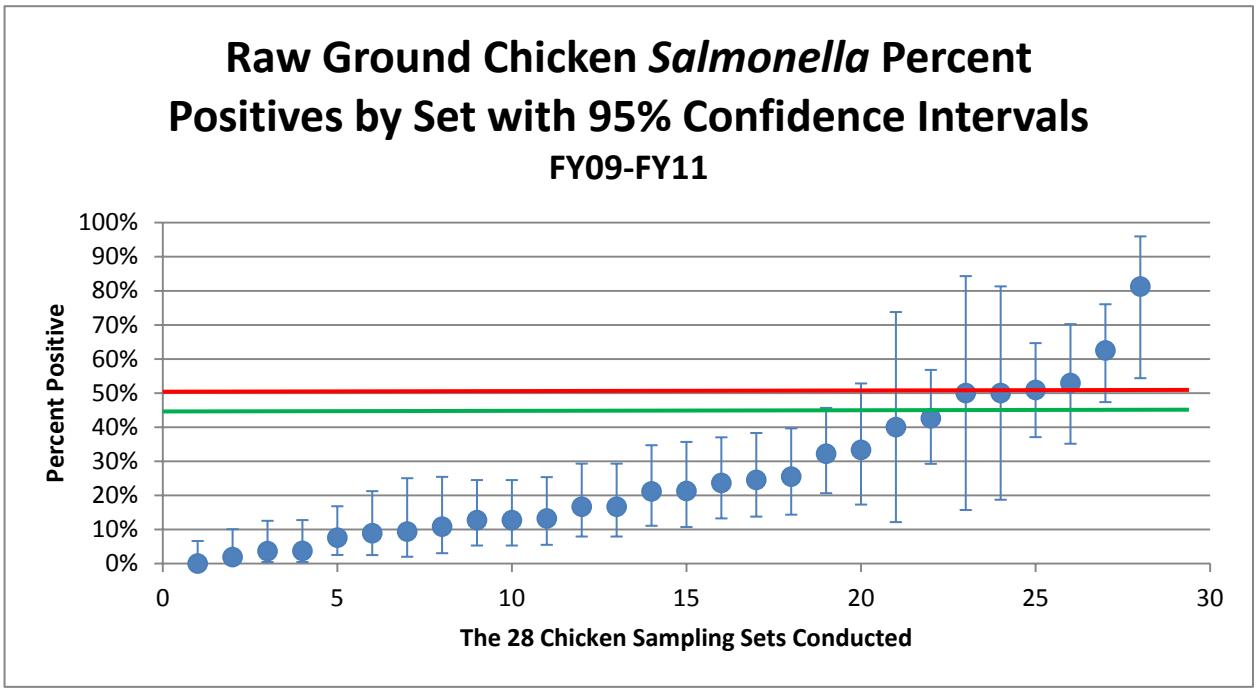

Figure 6: Raw Ground Chicken *Salmonella* Percent Positives by Set with 95% Confidence Intervals

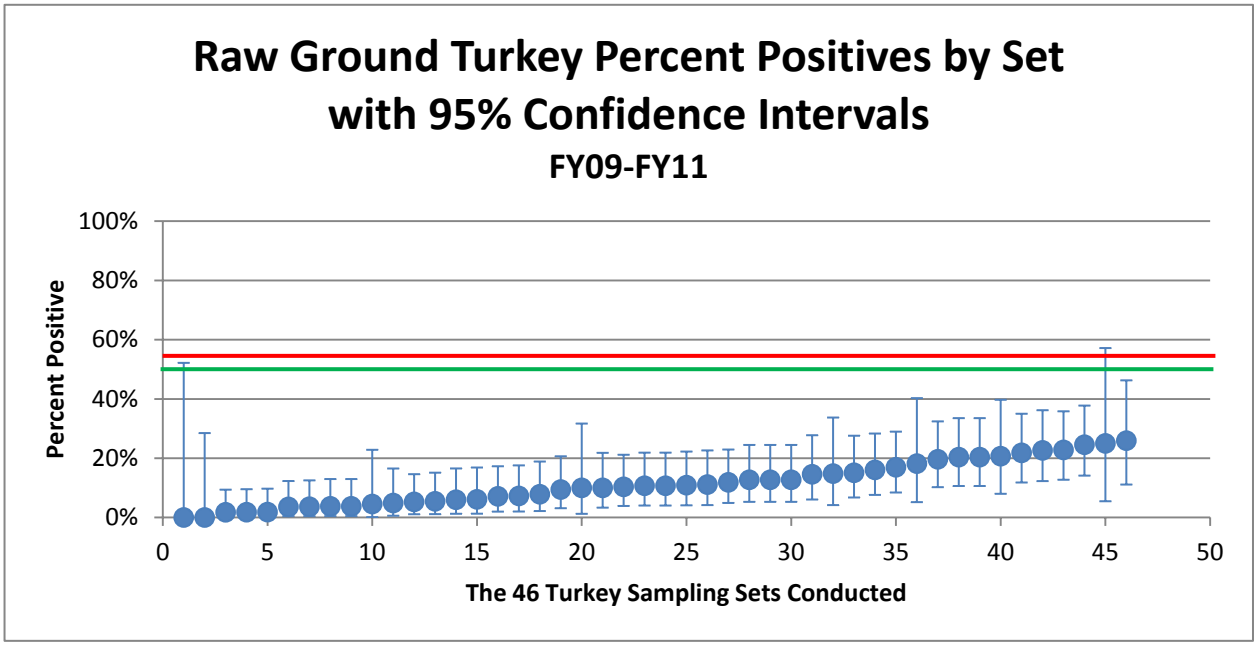

Figure 7: Raw Ground Turkey *Salmonella* Percent Positives by Set with 95% Confidence Intervals

Figure 6 and 7 show that most sets are well below the maximum threshold (represented by the red solid line) and the FSIS performance standard (represented by the green solid line) that was set in the mid-1990s. While no raw ground turkey sets were above either threshold, only six raw ground chicken sets were above both thresholds, and only one of those sets had 50 or more samples.

3.6 Time Independence within a Sampling Set

All raw ground poultry sample sets were broken into windows of ten samples. There were 17 sets that contained 50 or more samples for raw ground chicken and 34 sets that contained 50 or more samples for raw ground turkey. Figure 8 shows box plots for the sample windows of raw ground chicken sets, where the dashed line is the average. For all windows—except 41 to 50—the lower quartile is zero. For samples 41 to 50, the median is not visible because the lower quartile and the median are both 10%. Figure 9 shows the same for raw ground turkey. This time, all of the lower quartiles are zero. Also, the mean for the sample window 21 to 30 in Figure 9 is the same as the median (10%), so it is not visible.

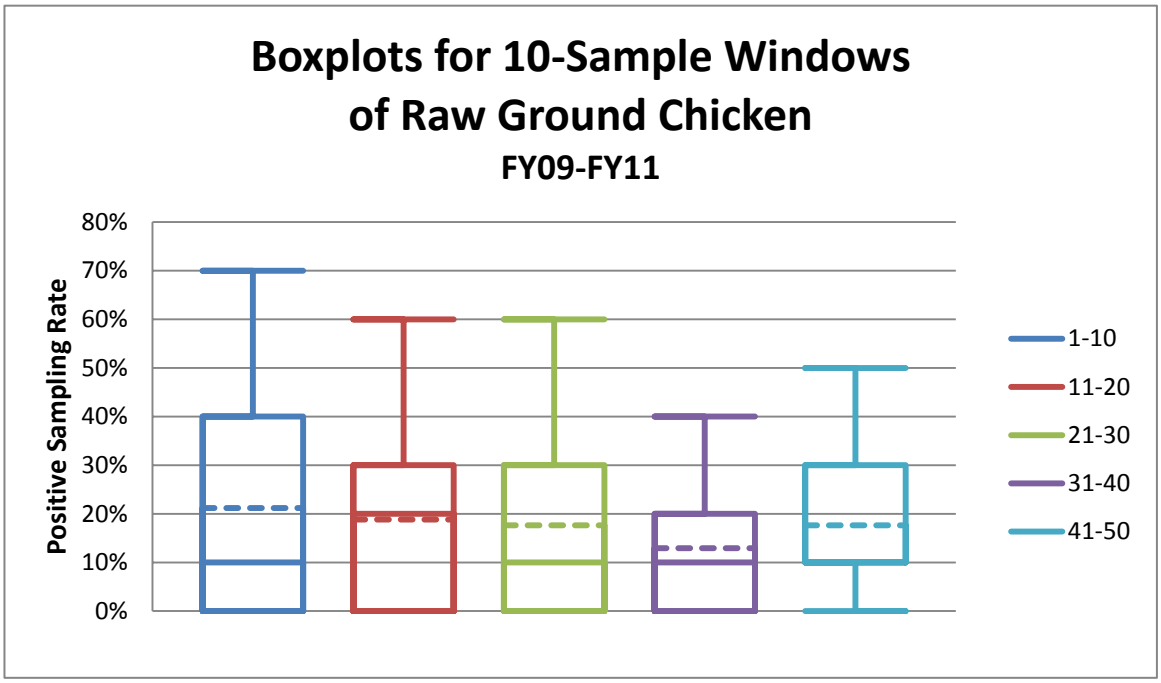

Figure 8: Box plots for 10-Sample Windows of Raw Ground Chicken for FY09-FY011.

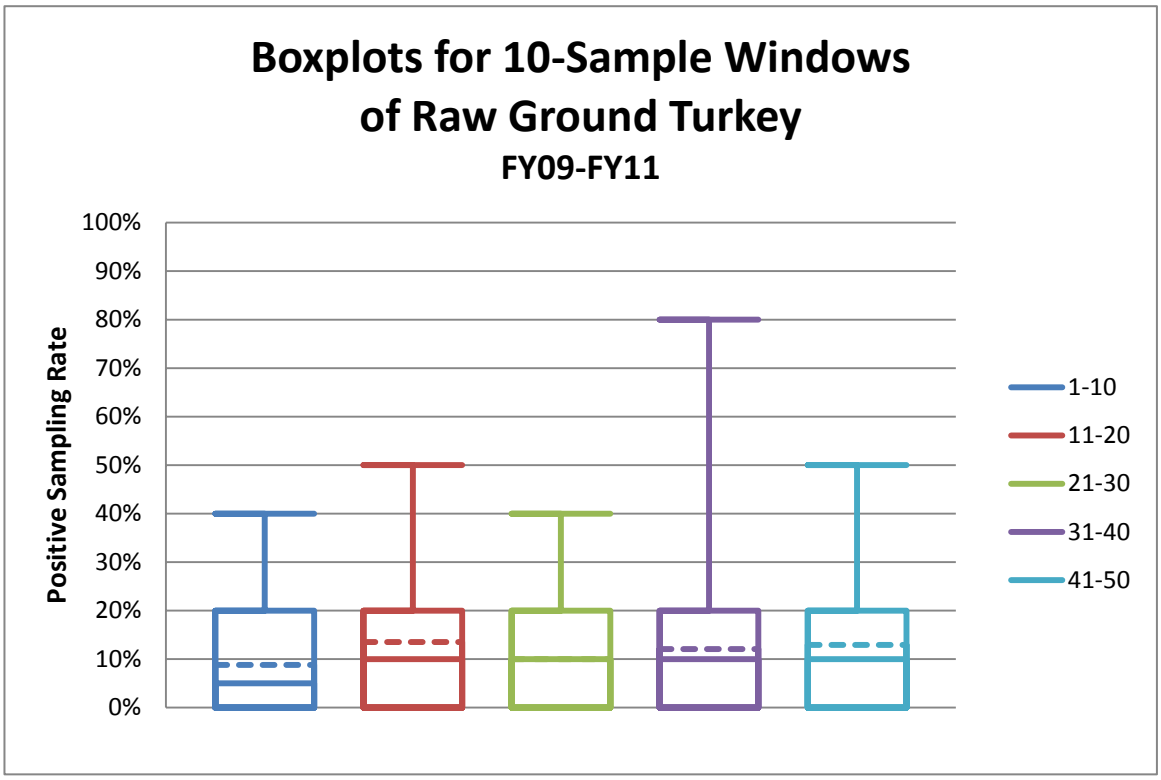

Figure 9: Box plots for 10-Sample Windows of Raw Ground Turkey for FY09-FY011.

Additionally, for both chicken and turkey, a single-factor analysis of variance (ANOVA) was performed. An ANOVA test is used by statisticians to see if a group of means are the same, which is the null hypothesis. If the p-value returned is below a set threshold (usually less than 0.05), the null hypothesis is rejected. The test checks to see if the average sampling rate in sample window 1-10 is the same as sample window 11-20, and so on. In both cases, the p-value was large (0.764 for chicken and 0.521 for turkey), indicating that the sampling windows have little impact on the positive sampling rate.

3.7 First and Second Halves of Sampling Sets

Since the FSIS standard set size for chicken and turkey is 53, windows of 25 samples were used to compare the first and second "halves" of a standard set. Table 4 shows some statistics on the difference between the second and first sets of 25 samples. A negative difference indicates that the positive sampling rate in the first 25 samples was higher than the rate in the second 25 samples. In other words, the positive sampling rate was lower in the second 25 samples than in the first 25 samples. The raw ground chicken sets had a slightly negative average difference, while the raw ground turkey sets had a slightly positive average difference. However, both chicken and turkey had sets with both positive and negative changes.

	Raw Ground Chicken	Raw Ground Turkey
Count	17	34
Average Difference Between Positive Sampling Rate in First and Second Half of Sets	-2.7%	0.5%
Standard Deviation	10.6%	12.2%
Minimum Difference	-25.3%	-20.0%
Median Difference	-0.7%	1.7%
Maximum Difference	13.3%	36.0%
Number of Sets with Negative Difference	9	14
Number of Sets with No Difference	1	3
Number of Sets with Positive Difference	7	17

Table 4: Description of the Difference between First and Second Groups of 25 Samples from Raw Ground Poultry Sets for FY09-FY11.

T-tests on both chicken and turkey sets were calculated to determine if the means of the first and second 25-sample windows were different. A t-test is similar to an ANOVA, except it only tests if two means are equivalent or not. Figure 10 shows the box plots for both raw ground chicken and turkey. Again, the p-values were large (0.617 for chicken and 0.826 for turkey), suggesting the difference in the means of the first and second halves were not statistically significant. This, combined with the analysis of the ten-sample windows, suggests that the positive *Salmonella* sampling rate within an FSIS sampling set is independent of time.

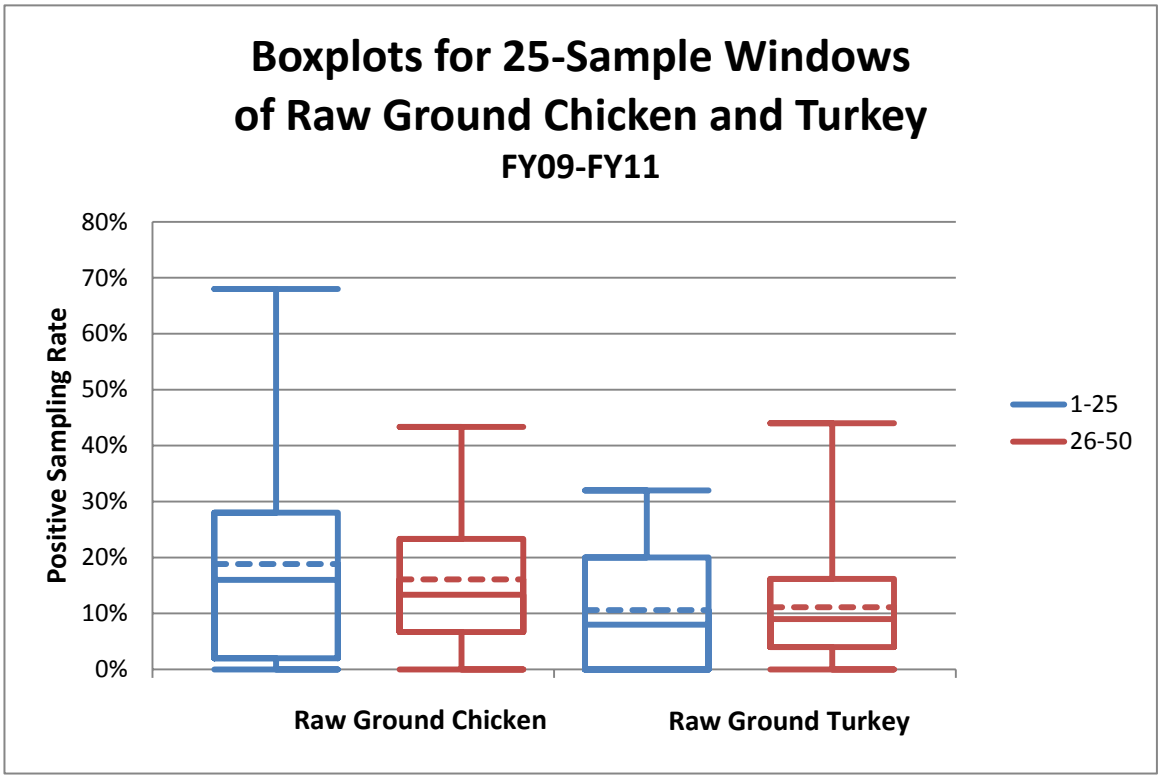

Figure 10: Box plots for 25-Sample Windows (Half of a Set) of Raw Ground Chicken and Turkey for FY09-FY011.

3.8 Meeting Performance Standards

Using the current FSIS performance standards, the failing threshold was calculated for different set sizes and then compared to the number of sets below these thresholds for the different set sizes. At 50 samples, 16 of 17 chicken sets were below the failing threshold of 25 positives. This is true at set sizes of 25, 30, and 40 as well. At 20 samples, only 15 of 17 chicken sets were below the failing threshold, which equates to ten samples. For turkey sets, 34 of 34 sets fell below the failing threshold at all sample sizes. Figures 11 shows the number of positives for both chicken and turkey sets at a sample size of 30. As indicated in the figures, the red line is the maximum number of positives allowed to meet the performance standard. The ordering of the sets in both figures is temporal from left to right.

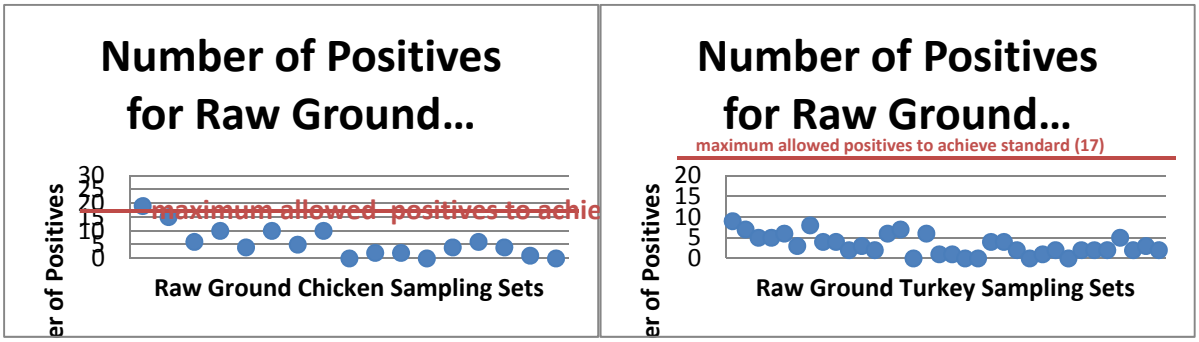

Figure 11: Number of Positives for Raw Ground Chicken and Turkey Sets at 30 Samples in FY09-FY11.

3.9 Falling Below Current Failing Thresholds

To be 95% confident that an establishment has passed a sample set, a one-sided confidence interval—an upper confidence bound—is needed. Table 5 shows the observed positive sampling rate below which there would be 95% confidence that the set's positive rate was below the failing threshold for both chicken and turkey .

N	Chicken	Turkey
20	28%	33%
25	30%	36%
30	32%	38%
40	34%	40%
50	36%	42%

Table 5: Observed Positive Rates to Ensure 95% Confidence of Passing Chicken and Turkey Sets at Sample Sizes N.

The confidence bound for a sample size of ten was not evaluated because it is substantially larger than the interval at a sample size of 20.

Figures 13 and 14 show the upper confidence bounds for raw ground chicken and turkey sets. The sets are in ascending order by positive sampling rate, and the red line marks the current failing threshold of 49.1%. The number of sets with upper bounds below the failing threshold for each of the set sizes in Table 5 was counted, but only some charts are included in this report for the sake of brevity. At 20, 25, and 30 samples, 12 of 17 chicken sets fell below the threshold. At 40 samples, there were 15 of 17 sets below the threshold, but at 50 samples, there were 14 of 17 sets that had upper bounds below the failing threshold. This means there was 95% confidence that the true positive sampling rate for these sets was below the failing threshold. If the sampling had halted at 40 samples, one set would have passed with 95% confidence that should not have. This demonstrates the uncertainty of confidence intervals; there is still a 5% chance that the true value will be above the 95% upper confidence bound.

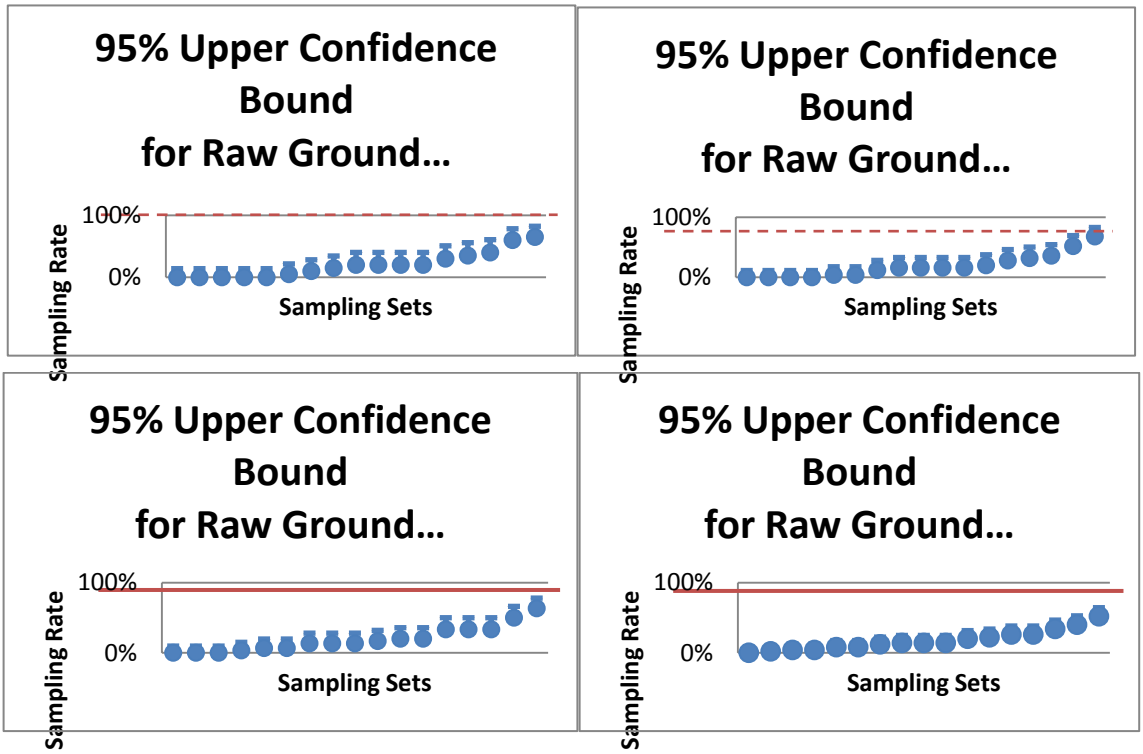

Figure 12: 95% Upper Confidence Bound for Raw Ground Chicken Sets at 20, 25, 30, & 50 Samples in FY09-FY11.

Figure 13 shows the upper bounds for raw ground turkey sets. At 20 and 25 samples, 33 of 34 sets fell below this threshold. At 30 samples, all sets fell below the threshold, and this remained true for both 40 and 50 samples.

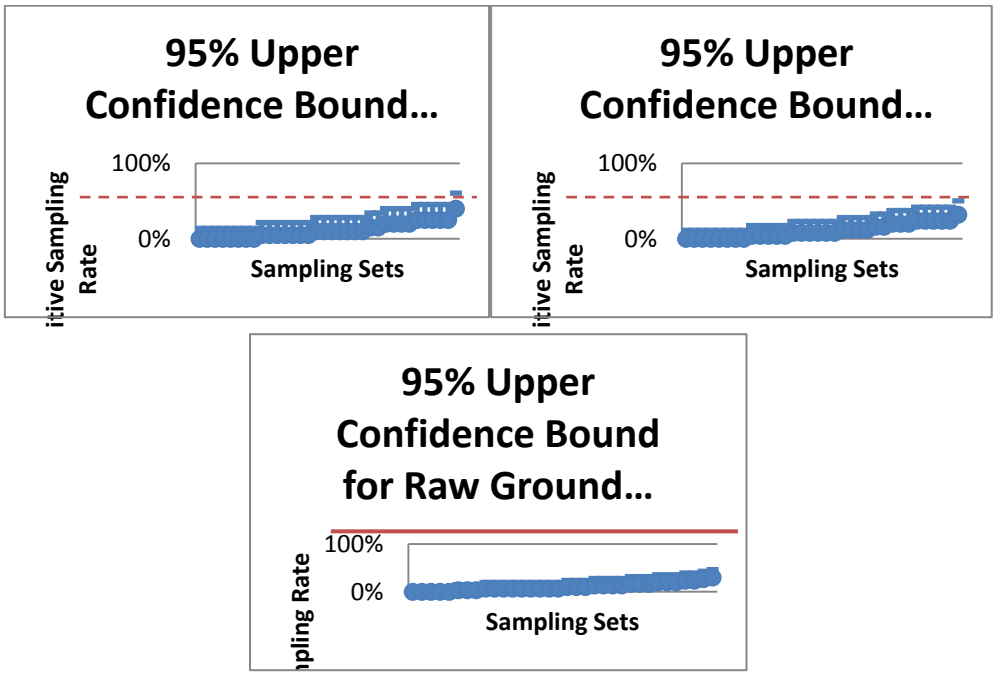

Figure 13: 95% Upper Confidence Bound for Raw Ground Turkey Sets at 20, 25, & 30 Samples in FY09-FY11.

Considering each of the sample sizes—and counting both chicken and turkey sets—yields the following. There was 95% confidence that 88% (45/51) of sets passed after 20 and 25 samples. After 30 samples, this confidence was obtained for 90% (46/51) of sets. After 40 samples, there was 95% confidence that 96% (49/51) of sets passed, but after 50 samples, this decreased to 94% (48/51) of sets.

Overall, this suggests that going as low as 20 samples could be effective at accurately assessing whether a set passes or not. However, at 20 and 25 samples, the confidence interval remains relatively long. Increasing the set size by five samples decreases the width of the confidence interval by 2%. After 30 samples, increases of ten samples are required for the same reduction, illustrating again the diminishing return on increasing sample size. Therefore, a sample size of 30 samples appears sufficient to ensure that a set has passed and balances precision with the desire to reduce the size of sets. Additional figures are included below.

3.10 Bounding the Change in Sampling as Set Size Increases

In this analysis, the change in the positive sampling rate as the set size increases was bounded. To do so, the change between 30 and 50 samples was evaluated, as well as between 40 and 50 samples. This method is less precise than using the confidence intervals discussed above. However, it provides a means of empirically demonstrating similar results to those achieved using confidence intervals. Figure 14 is a histogram of the difference between 30 and 50 samples for raw ground chicken sets. The average change was -1.6%, with a standard deviation of 5.5%. Figure is a histogram of the change from 40 to 50 samples for raw ground chicken. Here, the average was 0% and the standard deviation 1.8%.

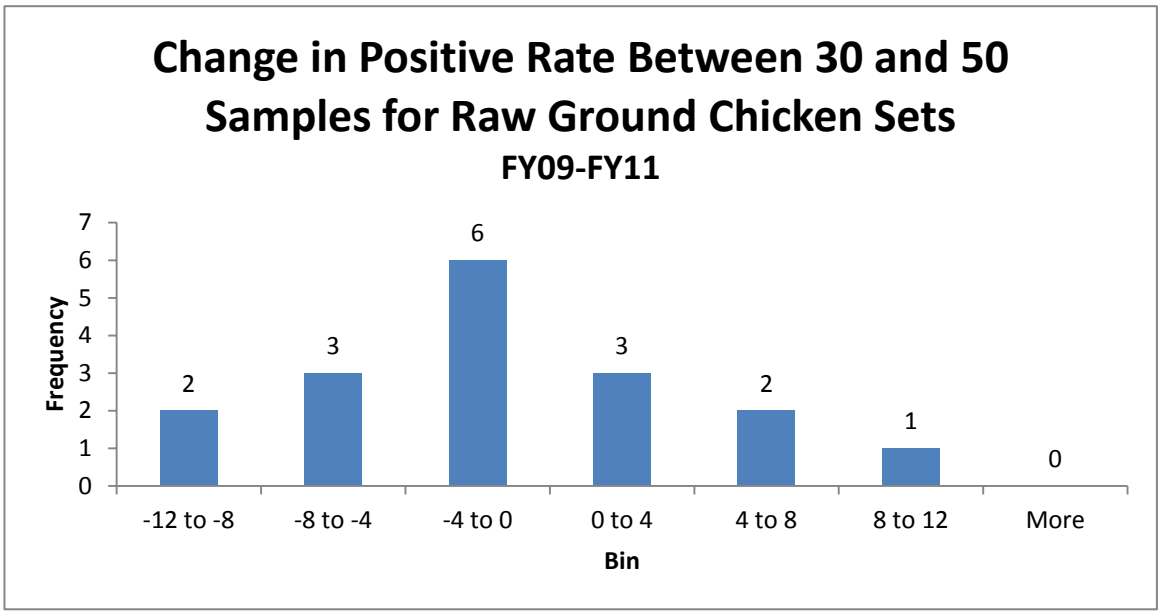

Figure 14: Histogram of Changes in Positive Rate Between 30 and 50 Samples for Raw Ground Chicken Set in FY09-FY11.

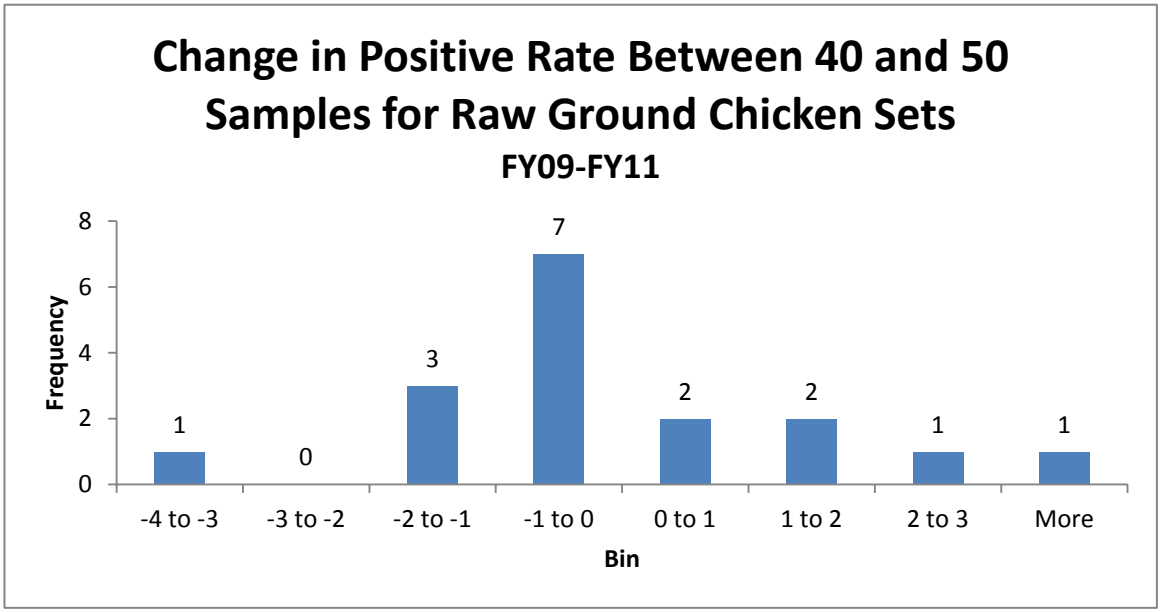

Figure 15: Histogram of Changes in Positive Rate Between 40 and 50 Samples for Raw Ground Chicken Set in FY09-FY11.

The sets of raw ground turkey demonstrated a similar behavior. Figures 14 and 15 are histograms for the change from 30 to 50 samples and 40 to 50 samples, respectively. The average difference between 30 and 50 samples was 0.7%, with a standard deviation of 5.5%. The average difference between 40 and 50 samples was 0.4%, and the standard deviation was 2.6%.

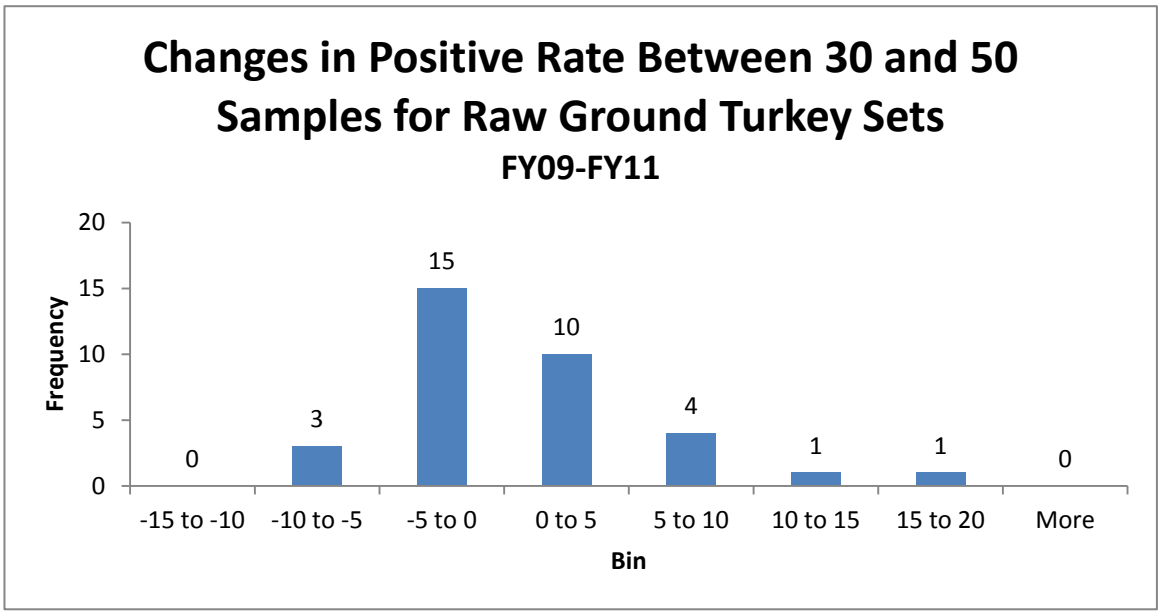

Figure 16: Histogram of Changes in Positive Rate between 30 and 50 Samples for Raw Ground Turkey Set in FY09-FY11.

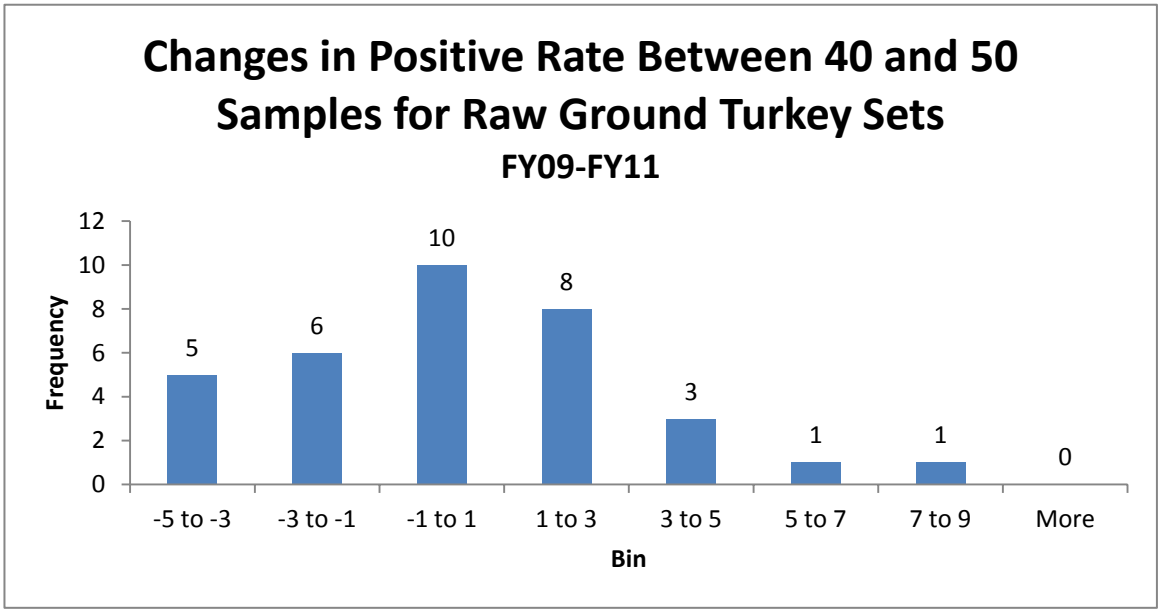

Figure 17: Histogram of Changes in Positive Rate Between 40 and 50 Samples for Raw Ground Turkey Set in FY09-FY11.

Each of the histograms in Figures 14 through 17 had an approximately normal shape. The Shapiro-Wilk test for normality tests the null hypothesis that data come from a normal distribution. A test for each set of changes failed to reject this null hypothesis. This suggests that the changes are normally distributed.

Within a normal distribution, 95% of all observations fall within two standard deviations of the mean. This means that—for both chicken and turkey—a 95% confidence interval can be

provided on the change that would occur with increasing the sample size from 30 to 50 samples and from 40 to 50 samples. Table 6 shows the lower and upper bounds for these changes.

	Lower Bound	Upper Bound
Chicken		
30 to 50	**-12.6%**	**9.5%**
40 to 50	**-3.6%**	**3.6%**
Turkey		
30 to 50	**-10.4%**	**11.7%**
40 to 50	**-4.8%**	**5.5%**

Table 6: 95% Confidence Interval on Change in Positive Rate Between 30 to 50 and 40 to 50 Samples FY09-FY11.

As an example, the upper bound on change from 30 to 50 samples (for chicken) was 9.5%. This means that FSIS can be 95% confident that increasing the set size to 50 samples will not cause the set to exceed the failing threshold when the sets are at or below 39.6% positive with 30 samples. At these bounds, FSIS has 95% confidence that 96% (49/51) of sets passed at 30 samples. This reaffirms MITRE's recommendation that 30 samples is a sufficient set size.

The change from 20 to 50 samples was also tested. However, as mentioned at the beginning of this section, bounding the change is less precise than using binomial confidence intervals for the observed positive sampling rate. Therefore, the change between 20 and 50 samples is excluded from discussion in this report.