U.S. Department of Health and Human Services

Office of Inspector General

Statistical Sampling: A Toolkit for MFCUs

September 2018

An OIG Toolkit

OIG-12-18-1

The purpose of this toolkit is to outline the basics of statistical sampling for use by State Medicaid Fraud Control Units (MFCUs) in calculating improper payment amounts.

Introduction

The purpose of this toolkit is to serve as a practical training guide to help State Medicaid Fraud Control Units (MFCUs) design effective statistical samples. The guide is not intended to establish formal standards or to restrict the methods available to the MFCUs to complete their mission. Because statistics is a broad field with a wide variety of effective and valid methods, a sample or estimate may be valid even if it does not follow the steps in this guide.

Statistical sampling is a widely accepted methodology in an audit or other review of healthcare claims to identify improper payment amounts. When performed in a valid scientific manner, statistical sampling permits the Government to estimate overpayments in a universe of claims that may be too voluminous or complex to permit a claim-by-claim review. Statistical sampling saves the time and expense of reviewing the entire universe of claims by allowing a small sample of claims to be analyzed.

Statistical sampling involves selecting a subset of items from a larger group (often referred to as the "frame" or "population") and using the results of the sample to estimate a characteristic of that larger group. For example, sampling could be used to identify the total amount of improper claims made by a healthcare provider that submitted hundreds or thousands of claims, without examining each of the individual claims. Sampling avoids the cost and practical challenge of examining a large number of claims.

This toolkit is not intended to provide legal guidance regarding the use of statistical sampling. Significant case law supports the use of statistical sampling to calculate overpayment amounts and for other purposes. However, it is important to research and identify any jurisdiction-specific limitations on sampling that might apply to a case. More generally, MFCUs should ensure that sampling is performed correctly and should be prepared to defend the sampling methodology in court or other legal forum.

Valid samples and estimates can be calculated using a variety of statistical packages. The examples in this toolkit refer to the RAT-STATS software package that is maintained by the Office of Inspector General (OIG), U.S. Department of Health and Human Services (HHS), and is available free of charge on OIG's website here. RAT-STATS can be used to determine a sample size, generate random numbers, and calculate statistical estimates. However, RAT-STATS is only one of several statistical packages available. Several commercial products that provide similar services are also commonly used.

Need Additional Help?

This toolkit was prepared by HHS OIG for the MFCUs. For general questions about the sampling process, or the operation of the RAT-STATS software, please contact Douglas.Bronson@oig.hhs.gov. MFCU staff who need detailed assistance should contact a statistician within their own State agency or contractor.

Contents of the Toolkit

Statistical Sampling Step-by-Step

Step	Description	Tips
1. Define the objective (Example)	Identify whether sampling is needed to identify a potential improper payment.	If planning to seek assistance from a statistician, get the expert involved early in the sample planning process.
2. Identify the target population to be sampled (Example)	Identify the provider, time frame, and claim types (when relevant) that represent the target of the review. This is the target population. When defining the time period, consider whether there were any changes that might affect the sample review (changes in ownership at providers, changes in relevant rules and regulations, etc.).	Statistical methods can be applied even with very large datasets, so do not worry about the target population being too big as long as it is composed of sample units that are relevant to the review.
3. Identify the method of measurement (Example)	Select the methods that will identify the improper payment amount for each item included in the sample. The focus here should not be on statistical methods but on the criteria to be applied (e.g., relevant coverage or payment rules) and the tests to be performed (e.g., medical review).	The decision on the measurement approach may have implications on the choice of sample unit.
4. Identify the sample unit (Example)	Given the target population and the method of measurement, decide what represents a single unit for the purpose of the review. For example, if the review involves inspecting claims to see whether the underlying service is medically necessary, then a claim line might be a reasonable sample unit. Other examples of potential sampling units include a full claim, a beneficiary, and a date of service.	The goal should be to select a sample unit that is consistent with the method of measurement; however, keep in mind that the greater the difference in overpayment amounts between items in the sampling frame, the less precise the resulting point estimate will be (i.e., the larger the margin of error).
5. Identify the sampling frame (Example)	To pull the sample, first compile a list of items (hardcopy or electronic) that compose the population. Think of this file (referred to as the sampling frame) as the dataset that would be used for a 100-percent review. Within the sampling frame, each row should relate to a single sample item. For example, if a beneficiary is the sample unit, then the frame would be a list of beneficiaries. The sampling frame need only include items that are related to the objective. One test of whether to remove an item from the frame is to consider whether the item would be included as part of a 100-percent review. If the answer is yes, then consider keeping it in the sampling frame.	Make sure the sampling frame is free of duplicates and other anomalies. For example, be sure any data files combined across managed care organizations are consistent. Save a copy of the sampling frame in case the sample has to be replicated. Document in a clear, detailed fashion how the sampling frame was constructed. Number the sampling frame so that the order of the frame when the sample was selected can be replicated. When constructing the sampling frame, consider removing $0 paid claims. Also, it is often advisable to use only final action claims.

Step	Description	Tips
6. Decide on a sample design (Example)	The most straightforward design is a simple random sample in which items are selected at random from the sampling frame. Another approach is to use stratification. Stratification involves dividing the frame into separate groupings and then pulling a portion of the overall sample from each grouping—for example, dividing a frame in half with the higher dollar items in one half and the lower dollar items in the other half and then selecting items at random from each half.	Seek advice from a statistician before trying a particular sampling approach for the first time. A poorly constructed stratified random sample will perform worse than a simple random sample. For more details on stratification, please refer to Appendix G.
7. Decide on a confidence level (Example)	Choose the confidence level for the calculation of the estimate. The confidence level is usually determined by agency policy and is decided on before pulling the sample. The confidence level represents how often the upper and lower limits of an estimate should contain the population quantity of interest (e.g., the overall improper payment amount). The upper and lower limits together are known as the confidence interval.	Most agencies, as a matter of policy, use either a 90- or 95-percent confidence level. A 90-percent confidence level is common in the Medicare administrative appeals process. A 95-percent confidence level is more common in academic settings. As the confidence level of an estimate increases, the associated confidence interval gets wider.
8. Decide on the sample size (Example)	The choice of sample size involves a tradeoff between the time or cost required to review the sample and the precision of the estimate (i.e., how close on average it will be to the target population average of interest). The more variability in the sampling frame, the larger the sample size needed for any given precision amount.	Consider working with a statistician to develop a policy to guide the choice of a sample size for the review. For additional information about the choice of sample size, please refer to Appendix G.
9. Document the sample design (Example)	Draft a sampling plan that describes the sampling frame, sample unit, sample design, sample size, sampling method, and planned estimation approach. Appendix C contains a list of documents that are often requested by defense attorneys.	Consider defining a formal review process for plan clearance, including review by a statistician or a person with equivalent expertise in probability sampling and estimation methods.
10. Generate the random numbers (Example)	Ensure that each record in the sampling frame is uniquely and consecutively numbered. Use a valid random number generator to generate the random numbers for the sample.	To ensure that the sample can be replicated, save the random seed value that was used to generate the random numbers, along with the random numbers themselves.
11. Select the sample (Example)	Identify the row numbers in the sampling frame with unique numbers that match the random numbers generated in the previous step. Some programs generate random numbers and select a sample in a single step.	Save a copy of the sampling frame that was used to pull the sample. Sample selection can be done manually or through an automated function.
12. Review the sampled items (Example)	Review each sampled item. No items should be excluded from the sample or replaced without consultation with a statistician.	Be sure to save a copy of the sample results. Items reviewed outside of the sample cannot be included as part of the estimate calculations.
13. Calculate the statistical estimate (Example)	Once each sample item has been reviewed, use a valid statistical program to estimate the target frame quantity (e.g., the total improper payment amount).	*Do not attempt to calculate statistical estimates by hand*. Consider having a statistician review the estimate methodology.

Appendix A: Application of Statistical Sampling Steps to a Case Involving Potential False Billing by a Physician

1. Define the objective (Step Description)

With the assistance of a data analysis group, OIG investigators identified a physician who was billing for a significant, separately identifiable evaluation and management service (modifier 25) in more than 96 percent of claims. The investigators were interested in identifying any overpayments made to this provider because of inappropriate billing of claims with modifier 25.

2. Identify the target population to be sampled (Step Description)

The investigators decided to restrict their review to a 3-year period (2014 through 2016). The population was all claims with modifier 25 paid by the target physician during that time.

3. Identify the method of measurement (Step Description)

The investigators planned to subpoena medical records from the physician and then have a qualified medical coder perform a review to identify whether the records supported the codes billed on the claims submitted.

4. Identify the sample unit (Step Description)

The investigators could have defined the sampling unit as a claim or a beneficiary. The investigators decided to use the claim as the sample unit. The reasons for this choice were that the investigators needed to review individual claims as part of their planned investigative procedures, and the error amounts obtained from their reviews were likely to vary less across claims than beneficiaries.

5. Identify the sampling frame (Step Description)

The investigators obtained a database that contained all of the claims for the physician with the target modifier that were paid between 2014 and 2016. The investigators filtered this dataset to remove claims that had been canceled and claims for which no funds had been paid to the physician. After applying these filters to the data, a total of 2,100 claims remained. These claims were placed in a separate file and numbered from 1 to 2,100. This numbered file was the sampling frame.

6. Decide on a sample design (Step Description)

The investigators decided to use a simple random sample. This design is the easiest to perform and is also easy to explain and defend. If there had been bigger differences between the claims in the frame, then using a stratified sample may have made more sense. If the investigators needed separate estimates for different parts of the frame, then they might have used multiple simple random samples.

7. Decide on a confidence level (Step Description)

Following their agency's policy, the investigators used a 90-percent confidence level.

8. Decide on the sample size (Step Description)

The choice of sample size involves a tradeoff between the resources and time required to complete a review of the sample and the precision of the resulting statistical estimate. In this case the team decided on a sample size of 50. (See Appendix G for a more detailed description of the factors to consider when deciding on the sample size.) The sample was selected based on a cost-benefit analysis where the team balanced the time and resources required to review the sample against the improved precision that would result from the larger sample size.

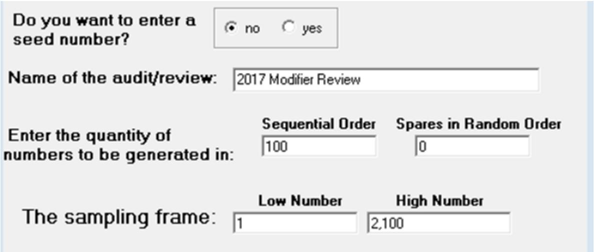
9. Document the sample design (Step Description)

The investigators drafted a three-page sampling plan that was signed by a supervisor, an attorney assisting on the case, and an auditor who had experience with sampling. The document described the objectives of the investigation, the target population, the sampling frame, the sampling unit, the sample design, the sample size, the source of random numbers for the sample, the method used to select the sample, the seed number, the characteristic(s) to be measured, the planned estimation approach, and the source of the data used to construct the sampling frame.

10. Generate the random numbers (Step Description)

An individual who had experience with sampling generated 50 random numbers using RAT-STATS. Figure 1, shows the parameters that were used to generate the sample.[1]

Figure 1



Do you want to enter a seed number? ● no ○ yes

Name of the audit/review: 2017 Modifier Review

Enter the quantity of numbers to be generated in: Sequential Order 100 Spares in Random Order 0

The sampling frame: Low Number 1 High Number 2,100

11. Select the sample (Step Description)

An investigator identified the line numbers in the frame that matched the random numbers generated by RAT-STATS. For example, the first random number was 18, which corresponded to the 18th record in the sampling frame (i.e., the record in the sampling frame with a line number of 18). Rather than match the random numbers to the sampling frame by hand, the investigator could have performed this step using automated software. The selected items were carefully reviewed by a second individual to ensure that the selected claims aligned with the random numbers generated from RAT-STATS.

12. Review the sampled items (Step Description)

The investigators subpoenaed the medical records associated with the claims selected in the sample. Once the investigators received the subpoenaed records, the investigators gave the records to a trained medical coder who reviewed each sampled item and identified whether the claimed amounts were

[1] When RAT-STATS outputs random numbers, it also outputs the seed used to generate those numbers. The seed is generated automatically by RAT-STATS when not entered by the user. The seed can then be used by anyone with the RAT-STATS software to re-create the random sample. This type of feature is standard in many statistical packages.

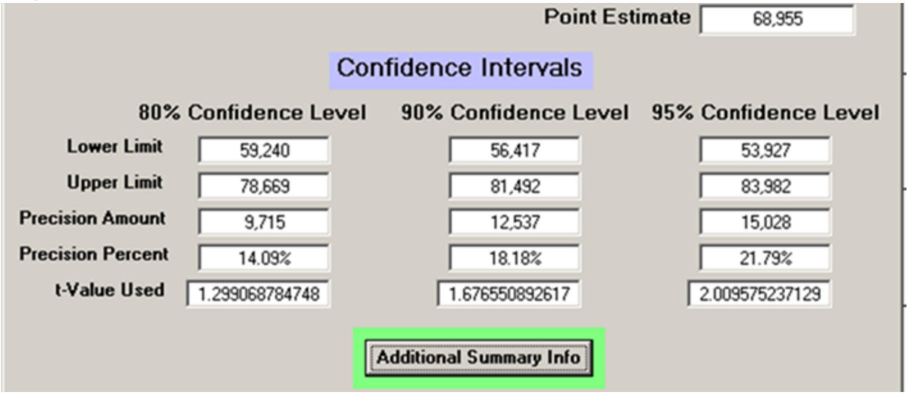
allowable. When the coder determined that the claimed amounts were not allowable, the investigator identified the amount of the resulting overpayment. The investigator created a spreadsheet that contained the total amount overpaid to the provider for each sampled item. If no overpayment was identified, the item was coded as having no overpayment. All 50 sampled items were included in the results file.

13. Calculate the statistical estimate (Step Description)

The spreadsheet created in Step 12 was entered into the RAT-STATS unrestricted variable appraisal module. For a simple random sample, the spreadsheet should include two columns of information. The first column should contain the sample number (or other identifier) and the second column the overpayment amount. The first column is ignored by RAT-STATS but is useful for anyone reviewing the sample results file.

Figure 2 contains example output from RAT-STATS. Three numbers in this screenshot are of greatest interest. The first number is the point estimate. The point estimate represents the best approximation of the overpayment amount ($68,955 in this case). Because the investigators used sampling, the actual overpayment amount in the frame may be larger or smaller than this estimate. Additional information is needed to capture how precise the estimate actually is. To measure the precision, refer to the confidence interval. Recall that the investigators in this case had decided to use a 90-percent confidence level. The interval associated with this confidence level ranges from $56,417 to $81,492.

Figure 2



Point Estimate 68,955

Confidence Intervals

	80% Confidence Level	90% Confidence Level	95% Confidence Level
Lower Limit	59,240	56,417	53,927
Upper Limit	78,669	81,492	83,982
Precision Amount	9,715	12,537	15,028
Precision Percent	14.09%	18.18%	21.79%
t-Value Used	1.299068784748	1.676550892617	2.009575237129

Figure 2 Note: The confidence level, lower limit, upper limit, and precision amount are all explained in Appendix E. The percent precision is the precision amount divided by the point estimate. The t-value is a technical number that is used as part of the calculation of lower and upper limit.

Taken together, the estimate for the overpayment is $68,955, and the 90-percent confidence interval for the overpayment ranges from $56,417 to $81,492.

Epilogue: The Department of Justice successfully used the statistical estimate during negotiations, which resulted in a settlement.

Appendix B: Example Sampling Plan From Audit A-07-16-03209

The purpose of this section is to provide a basic example of a sampling plan, a version of which appeared in an actual OIG audit report. The sections here are not exhaustive of the type of information that could be included in a sampling plan. For example, a plan could describe the reasoning behind the choice of sample unit, sample design, and sample size.

TARGET POPULATION

The target population consisted of unique non-emergency transportation (NET) paid claims with positive Medicaid reimbursements for NET services that the State Medicaid agency provided and paid for during fiscal years (FYs) 2012 through 2014 (October 1, 2011, through September 30, 2014).

SAMPLING FRAME

The sampling frame consisted of 623,396 unique NET paid claims with procedure codes A0120, A0130, A0140, S0209, S0215, T1001, and T2003 in the 15 counties with the highest total payments for NET services during the audit period. The reimbursement amount associated with these 623,396 claims totaled $17,518,245 ($9,765,329 Federal share) for this period. The procedure codes represented all NET services except for (1) ambulance services and (2) services for which the vehicle was provided by an individual (e.g., a family member or neighbor).

SAMPLE UNIT

The sample unit was one NET paid claim.

SAMPLE DESIGN

The auditors used a simple random sample.

SAMPLE SIZE

The auditors selected 100 unique NET paid claims.

SOURCE OF RANDOM NUMBERS

The auditors generated the random numbers with the OIG, Office of Audit Services, RAT-STATS.

METHOD OF SELECTING SAMPLE UNITS

The auditors consecutively numbered the sample units in the sampling frame from 1 to 623,396. After generating 100 random numbers, the auditors selected the corresponding frame items.

ESTIMATION METHODOLOGY

The auditors used RAT-STATS to estimate the amount of the unallowable payments for NET services and to estimate the financial impact of the deficiencies associated with driver and vehicle maintenance checks. The confidence interval for this estimate was calculated at the 90-percent confidence level.

Appendix C: Commonly Requested Sample Documents

(See the Sample Documentation Step in the Step-by-Step List)

The list below identifies documents that may be requested by outside parties, including defense attorneys.

- The sampling plan or similar document that includes a description of the:
  - sampling frame,
  - sample unit,
  - sample design,
  - sample size, and
  - population quantities to estimate.
- File(s) describing any steps taken to create the sampling frame.
- File(s) describing any steps taken to verify the reliability of the sampling frame.
- The random seed that was used to generate the random numbers for the sample. Without the seed number, it is not possible to recreate the random sample.
- The output of the program used to generate the random numbers for the sample.
- The numbered frame used to pull the statistical sample.
- File(s) with the overpayment amount for each sampled item.
- File(s) with the output from the valid statistical software program used to analyze the sample results.
- File(s) describing any communications with technical and subject matter experts about the sample.
- The original data extract that was used to create the sampling frame. This information is often requested regardless of whether it is actually needed to validate the sampling frame.

Appendix D: Common Misconceptions About Statistics

Myth 1: Very large samples are required to handle very large frames.

Reality: Unless it is very small, the size of the frame has limited impact on the required sample size. Consider a sample of 100 items. If that sample has a precision of plus or minus 9.5 percent from a given frame of 1,000 items, then increasing the frame size to 100 trillion items would only worsen the precision by 0.5 percentage points to plus or minus 10 percent.

An equation for the impact of the frame size on the sample size required to reach a given precision can be found in Cochran's textbook on sampling.[2] Notably, Cochran first presents the sample-size calculation for a frame that is infinitely large and goes on to explain the adjustment for smaller frames in the example above.

Myth 2: It is important to check the sample to see if it is actually random.

Reality: A sample is random if it is generated from a valid random or pseudo-random number generator.[3] The principles of survey sampling laid out by Cochran and similar survey sampling authorities require that each sample item has a known probability of being selected. These probabilities cannot be identified if sample items or entire samples are thrown out because they are deemed to be not "random enough."

More generally, the fact that samples may be unusual is captured by the use of a confidence interval. Even though any given sample might not perfectly reflect the frame, the confidence interval will contain the true population value of interest for a large majority of samples. For example, a valid 95-percent confidence interval should include the true value of interest 95 percent of the time. The confidence interval is designed to account for the sample size, the frame size, the choice of sample unit, and all sources of variability in the sampling frame.

Myth 3: A sample is not valid if the sample items are related to each other.

Reality: Several statisticians have recently argued that estimates must be adjusted when multiple claims within the sample are related. For example, these statisticians challenge cases in which a sample of claims includes multiple claims from the same beneficiary.[4] In practice, the relevance of any dependencies among sample items hinges on the method used to calculate the statistical estimate. Dependencies between frame items must be accounted for when using model-based but not design-based methods.

Design-based methods used in RAT-STATS and similar programs depend on the randomness arising from the sampling process. These types of methods are common when performing finite population sampling, but they are uncommon in most other areas of statistics. When using design-based methods, two sample items are independent if the probability of selecting one item has no relationship to the probability of selecting the other item. Given this definition, two items can be independent for sampling purposes regardless of how related they are otherwise. This fact can be

[2] William G. Cochran. 1977. *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons, p. 78.

[3] The term *sample* refers to the complete set of items selected for review, not an individual sample item.

[4] As a general rule, sample items not selected cannot be included when calculating a statistical estimate. For example, if a sample includes two claims from a beneficiary with five claims, the three unselected claims cannot be used to calculate a statistical estimate, even if all five claims are fraudulent.

verified through simulation and an inspection of the proofs underlying design-based methods. When design-based methods are used, no adjustments are needed to account for the relationship between sample items unless they were used to select the sample.

In contrast, model-based estimates assume that the population itself was generated by a random process. Given this approach, the statistician constructs a model of the process that is thought to have generated the population and uses the model to calculate an estimate given the observed sample results. The dependencies in the data, such as those that might be found when there are multiple claims from the same individual, are important when using these methods. Below is a simple demonstration to better illustrate the difference between model-based and design-based methods.

Suppose there are two sampling frames. The first sampling frame contains two patients. One patient has 500 correctly billed claims valued at $200 each. The other patient has 1,000 claims that were each incorrectly billed for $200.[5] The second sampling frame contains the exact same error and correct amounts as the first, but each of the 1,500 claims is associated with a unique patient (i.e., rather than 2 patients, there are 1,500).

Under a model-based approach, the number of patients is important for modeling how the sampling frame was generated. For example, the model could account for the possibility that claims made on behalf of one patient are more likely to be similar than claims made on behalf of different patients. If the model was accurate, the additional information about the patients could improve the precision of the resulting estimate. With a design-based approach, the number of patients is only of interest if the patient identities were used to design the sample. In the case of a simple random sample of claims, the patient identities would not be relevant. As long as the frame error amounts were the same, it would make no difference if the sampling frame contained 2 patients or 1,500 patients.

Both model- and design-based methods are well accepted within the field of survey sampling. A complete discussion of the advantages and disadvantages of the two approaches is beyond the scope of this toolkit.

Myth 4: An imprecise sample is inherently invalid.

Reality: The less precise a sample, the more uncertainty there is about the calculated estimate; however, this limitation does not render an imprecise sample invalid. The precision of a sample can be measured using a variety of methods, including the standard error, confidence interval, and the margin of error. When the precision is poor, the uncertainty in the sample can often be managed through the use of alternate estimates such as the lower limit of a confidence interval. If a sample is sufficiently precise for the purposes of the analysis,[6] or if an alternative conservative estimate can be used, then the fact that a more precise design was possible does not imply a flaw in the actual design selected. For example, a simple random sample would be valid even if a more precise stratified design was possible.

[5] The demonstration would hold just as well if errors differed within the two patients, but to simplify the explanation the error amounts are defined as being the same.

[6] There is no bright-line statistical rule for how precise a sample needs to be to reasonably rely on the point estimate.

Appendix E: Tutorial on Basic Sampling Terms

Notes: (1) Those utilizing this toolkit in electronic format may press the Alt key and the left arrow key to return to the original link location. (2) The terms below are most easily understood if reviewed in the order listed.

Random Seed — Each random sample is associated with a random seed. The random seed for the sample is needed to recreate the sample. The random seed is a number output by a random number generator that can be used to recreate a sample. When the seed is entered into the random number generator, it will recreate the previous sample rather than provide a new set of random numbers.

Population — The population is the group of items (e.g., claims) that are of interest. For example, suppose a provider is believed to have been fraudulently billing Medicaid for certain tests that the provider did not perform from 2012 to 2016. The target population might be all payments made to the provider from 2012 to 2016 that include potentially fraudulent billings.

Sampling Frame — The database from which a sample is drawn is known as the Sampling Frame. The terms *sampling population* and *sampling frame* are roughly interchangeable. Any statistical estimates calculated from a sample apply only to the sampling frame. The *target population* of interest may differ from the *sampling population* actually covered by the sampling frame. For example, if the sample is pulled from a database of 1,000 claims, then the estimate would be restricted to a *sampling population* of 1,000 claims even if the *target population* covers 1,500 claims.

Universe — Seek clarification of this term if encountered because it can have multiple meanings. In RAT-STATS, the term refers to the sampling frame.

Population Quantity — The quantity to be estimated. For example, the population quantity might be the recoverable dollar amount in the sampling frame of 1,000 claims.

Point Estimate — A statistically valid estimate of the population quantity. Point estimates should be calculated using a robust statistical software package.

Confidence Interval — The point estimate may be higher or lower than the actual population quantity. The confidence interval accounts for this uncertainty and is represented by an upper and lower limit. Across multiple samples, the lower limit will tend to be less than the actual target quantity of interest and the upper limit will tend to be greater. The point estimate will tend to be closer to the population quantity than the upper or lower limit but lacks the assurance associated with the limits.

Confidence Level — The confidence level represents how often the confidence interval is expected to contain the population quantity. For example, if a sample is re-pulled many times, roughly 90 percent of the 90-percent confidence intervals calculated across these samples would contain the actual recoverable dollars in the sampling population. The confidence level applies only to the confidence interval; *it does not apply to the point estimate*.

Precision — The better the precision, the more likely the point estimate will be close to the population quantity. The wider the confidence interval, the worse the precision. The width of the confidence interval is one measure of precision. The confidence level is NOT a measure of precision. (See also U.S. General Accounting Office. 1992. *Using Statistical Sampling*. GAO/PEMD-10.1.6, pp. 48–51 (GAO sampling guide).)

Appendix F: Steps That Can Have Unintended Consequences

This section contains events that may seem unimportant but can actually have a significant effect on the sample design. Becoming familiar with these potential issues allows for better planned samples and the identification of situations in which it might be useful to seek outside help.

Changing the sample size after the sample has been selected to account for initial sample findings.

- Example: A team decides to pull a sample of 100 claims to estimate the amount overpaid to a provider. The team reviews the first 10 claims and finds that 9 of the claims were improperly paid. Given the high error rate, the team thinks that a sample of 100 might be overkill. As a result, they decide to change their sample size to 30.

- Issue: Almost all standard statistical approaches assume a fixed sample size. Changing the sample size after reviewing the results may, in some circumstances, render the estimate invalid. Contact a statistician if interested in learning more about methods that allow for changing the sample size during the course of the review.

Failing to confirm the reliability of the frame before selecting the sample.

- Example: A team is in a rush to pull a sample for an investigation. In the middle of the review, the team realizes that the frame contains several duplicate items. Each item in the sample appears only once within the sample, but some of the items have corresponding duplicates in the frame.

- Issue: If there are duplicate items in the frame, then each item in the frame does not have the same chance of being selected. Simply removing the duplicates from the frame after the sample has been pulled does not solve this issue unless the sample is scrapped and re-pulled. Otherwise, the team would need to contact a statistician to see if there is a way to use the estimate despite the duplicates. These potential issues can be avoided if the frame is reviewed for duplicates before selecting the sample. A careful review of the frame can also help reduce the number of items that are not of interest to the target investigation (e.g., canceled claims).

Removing items from the sample because they are not relevant to the objective.

- Example: A team is reviewing a random sample of 40 claims as part of work to identify potential bases for imposing civil monetary penalties. After starting the review, the team finds that the provider canceled and repaid five of the claims before the sample was selected. The team removes the 5 claims and uses the results for the 35 sample items to calculate a statistical estimate.

- Issue: The presence of canceled claims in the sample implies that there are more canceled claims in the sampling frame. Removing the claims from the sample will create a mismatch between the sampling frame and the sample. Instead of removing the claims, the team could contact a statistician who can calculate a valid estimate that fully accounts for the canceled claims in the sampling frame.

Not accounting for the sampling design when calculating a statistical estimate.

- Example: A team uses a stratified design in which 200 sample items are selected from high-, medium-, and low-dollar strata. After reviewing the sample items, the team takes the dollar error rate in the total sample and multiplies it by the total transaction amount in the frame.

- Issue: Because the team used a stratified design, they must account for this design when calculating the statistical estimate. The easiest way to account for the design is to use a statistical software package to calculate the estimate rather than trying to calculate the estimate by hand. The failure to account for the design will result in an invalid estimate.

Failing to keep sufficient records to replicate a statistical sample.

- Example: A team uses the "Rand()" function in Excel to generate random numbers for a statistical sample.

- Issue: The "Rand()" function does not include a random seed and thus makes it impossible to recreate the random numbers used to pull the sample. Other common problems include the failure to keep the original sampling frame and the failure to number the sampling frame or unambiguously document the sorted order. The failure to keep such documentation does not necessarily render a sample invalid, but it can make the resulting estimate more difficult to defend.

Allocating samples across strata in an inefficient manner.

- Example: A team reviewing durable medical equipment decides to stratify by item type. The three item types are similar in dollar value and are thought to have the same chance of being in error. The team creates three strata and selects 35 claims from each stratum. It turns out that one item type makes up 90 percent of the items in the frame.

- Issue: The team allocated only 35 claims to cover a category that made up almost all of the sampling frame. These types of unbalanced designs can result in worse precision than a simple random sample. Poor precision does not undermine the validity of a sample design, but can make it more difficult to base an overpayment on the point estimate rather than the lower limit. Complicating the matter further, a reasonable allocation for a dollar estimate may not match what is reasonable when estimating the number of items in error. For more details on stratification, refer to Appendix G.

Appendix G: Sample Size, Outliers, Missing Data, Spares, and Stratification

This section contains a general discussion of several more technical topics. The coverage here is only introductory. Readers interested in these subjects should consider contacting a statistician.

SAMPLE SIZE

Attribute Versus Variable Estimates

The terms *attribute estimate* and *variable estimate* are referenced in several places within the RAT-STATS software. Similar concepts arise in all major statistical software packages even if the terminology is not the same. Consequently, it is important to understand which term applies to the selected characteristics to be measured. The two terms are described briefly below.

Attribute Estimate

For an attribute estimate, each sample item is reviewed and placed into one of several predetermined categories. In many cases, the sample will involve just two categories (e.g., correct or incorrect), though the term applies to samples that fall into three or more categories as well. To avoid confusion, focus on how the sample is being coded rather than on what quantity is being estimated. The results of attribute samples are generally used to estimate the number or percent of items in the frame that fall into each target category.

Variable Estimate

A variable estimate involves the measuring of a number for each item in the sample. Dollar amounts are the most commonly calculated variable estimates in the healthcare domain.

Examples:

(A) Counting the number of errors greater than $500 in a sampling frame may sound like a variable estimate because it includes dollar values; however, it would be an attribute estimate because for each sample item the final determination is categorical (i.e., whether the item involves an error greater than $500).

(B) Measuring the total number of cavities incorrectly filled across all beneficiaries may sound like an attribute estimate because it does not involve dollar values. However, if the sample unit is a beneficiary, then it would be a variable estimate because we are measuring the number of incorrectly filled cavities for each beneficiary.

(C) Taking the example in (B) above and making the sample unit a single cavity procedure would result in an attribute estimate because each sample item could be placed in a category (i.e., correctly filled or incorrectly filled).

(D) If the goal is to identify the number of beneficiaries with any incorrectly filled cavities, then the beneficiary sample could also result in an attribute estimate because beneficiaries would be in one of two categories (i.e., all correctly filled cavities or at least one incorrectly filled cavity).

Sample Size: General (Sample Size Step Description)

The choice of sample size is a balance between the resources required to review additional sample items and the benefit of a more precise statistical estimate. There is no definitive cutoff after which a sample becomes valid. Instead, each additional sample adds incrementally to the cost and precision of the design. As the sample size increases, each additional sample item will result in a smaller improvement in the precision. The caveat is that smaller sample sizes (especially those of fewer than 30) may be more difficult to defend for technical reasons that are beyond the scope of this toolkit.

In many agencies, the choice of sample size is determined at least partially by policy. Some agencies go as far as to define a fixed minimum cutoff. For example, the OIG audit group requires additional approval for any statistical samples that include fewer than 100 items. The use of such cutoffs is by no means universal. Moreover, it is often possible to obtain valid results even with sample sizes that are much smaller. Consider working with a statistician to develop a sample-size policy that is reasonable for the needs of a particular organization.

If funds are recovered from a provider using the statistical lower limit, then smaller sample sizes will tend to result in lower recoveries. As a result, it is often possible to defend a lower limit even if the sample size is small and the precision is poor. The downside is that the recovery, on average, will be far less than if a larger sample had been pulled. Generally, the more uncertainty in the sampling process, the smaller the lower limit needs to be to help ensure a conservative result.

If the estimate is sufficiently precise, it may be possible to recover funds using the point estimate rather than the lower limit. The precision required to rely on the point estimate is a legal question that goes beyond the scope of this guide.

The choice of precision is the most important decision driving the size of the sample. It is common for individuals new to statistics to focus on the confidence level rather than the precision when deciding on the sample. This focus has several shortcomings. First, the confidence level tends to have a smaller impact on the sample size than the precision. Second, many organizations use the same confidence level for all estimates (e.g., 90 or 95 percent). Finally, the confidence level is chosen before pulling the sample and does not change afterward. In contrast, the precision may turn out to be unexpectedly better or worse than what was planned when designing the sample. For further reading on the topic of precision, see pages 48–51 of the GAO sampling guide.

Sample Size: Attribute

Selecting a sample size is easier when the goal is an attribute rather than a variable estimate. To calculate a sample size for an attribute estimate from a simple random sample, one needs the confidence level, the anticipated rate of occurrence, the frame size (referred to as the universe size in RAT-STATS), and the precision (referred to as the desired precision range in RAT-STATS). Each of these elements is described in more detail below.

Confidence Level

The confidence level represents how often the confidence interval is expected to contain the population quantity. For example, suppose a sample is re-pulled many times. Roughly 90 percent of the 90-percent confidence intervals that were calculated across these samples would contain the actual recoverable dollars in the sampling population. The confidence level only applies to the confidence interval; it does not apply to the point estimate.

Anticipated Rate of Occurrence

This represents the percentage of time the target event would be expected to occur. The percentage is calculated out of the total number of items in the frame. For example, in estimating the number of claims that were incorrectly billed, the user could consider the percentage of claims expected to be incorrectly billed within the sampling frame. When unsure, the conservative approach is to set the anticipated rate of occurrence at 50 percent. This choice will result in a larger sample size than any other possible rate.

Precision

This value is the most important for determining the sample size. It reflects how far away the upper and lower limits are expected to be from the point estimate. The worse the precision, the less meaningful the point estimate will be. In RAT-STATS, the attribute sample-size determination module requests the precision range. This range represents the full width of the resulting confidence interval. For example, if the error rate is 50 percent and confidence interval ranges from 40 to 60 percent, then the precision range would be 20 percent. This measure of precision differs from the RAT-STATS variable-sample-size determination module, which uses the half width precision (i.e., 10 percent in the above example). Other software packages may use other measures for precision, so always check the user documentation.

The precision is sometimes referenced in terms of the margin of error. The margin of error is often expressed as the distance of the upper and lower limit from the point estimate. For example, suppose the estimated total overpayment is $100,000 with a 90-percent confidence interval that ranges from $85,000 to $115,000. In this case the margin of error could be expressed as plus or minus $15,000 or equivalently as plus or minus 15 percent.

Understanding the concept of precision is critical to being able to make a meaningful choice about sample size. See pages 48–51 in the GAO sampling guide for more information on the topic.

Sample Size: Variable

Deciding on the sample size for a variable estimate requires much of the information needed for an attribute estimate with an added wrinkle: A reasonably reliable measure of the *variability of what is being measured* is required to calculate the sample size for a variable estimate. For example, a measure of the variability of the overpayment amounts is required to estimate the total amount overpaid. There are two notable ways to identify this variability.

The first approach is to use any results gathered before pulling the sample. The results may be from previous work or from a small test sample. Ideally, the results would be measuring the same quantity that will be measured as part of the statistical sample. Entering those results into the "Using a Probe Sample" module in RAT-STATS can identify the sample size. Many other software packages have similar features.

The second approach is to use information in the frame that is thought to be related to the target population quantity. The most common example in healthcare is when a full payment will be disallowed for any sample item that is in error. In these situations, the frame will contain a mix of $0 error amounts and error amounts that are equal to paid amounts. If this relationship holds, then it is possible to estimate the variability of the frame given the variability of the paid amounts in the sample along with the percentage of items expected to be incorrect. RAT-STATS is able to perform this calculation with the "Using Estimated Error Rate" module. If using this module, keep in mind that the "Total Amount" and "Standard Deviation" fields refer to the paid transactions amounts listed in the frame. The situation here differs from the probe sample module, which requires the standard deviation for the target population quantity.

Another option is to rely on the lower limit rather than the point estimate. The use of the lower limit generally protects against challenges that a sample is not sufficiently precise. No matter how imprecise a sample, the lower limit will be conservative on average. This fact does not mean that sample size is irrelevant when it comes to the use of the lower limit. The smaller the sample size, the less the lower limit will tend to be. In extreme cases, the lower limit may be negative even if the sampling frame contains a substantial overpayment to the provider. In addition, as previously noted, small sample sizes (especially those of fewer than 30) may be more difficult to defend for technical reasons that are beyond the scope of this toolkit.

Consult the user guide of the software package used to see what type of sample-size features are available. A statistical expert may provide additional methods for determining sample size beyond those listed here.

OUTLIERS

An outlier arises when an error amount for an item is significantly larger than the remaining items in the stratum (or the frame if the sample is not stratified). Outliers are not generally a concern if the dollar values in the frame are similar or if the frame has been stratified in a reasonable manner.

When outliers[7] are present and not handled through stratification, they have two primary effects: (1) outliers worsen the precision of the extrapolation and (2) outliers tend to make the lower limit overly conservative (i.e., the lower limit is likely to be lower than it needs to be to reach the target confidence level). These two issues often arise together. A common alternative to using an imprecise point estimate is to use the lower limit. However, the lower limit is likely to be very small or even negative in a sample with an outlier. When these two effects are taken together, the presence of an outlier in a sample can undermine an otherwise well-designed sample.

The primary method for handling outliers in survey sampling is effective stratification. Another solution is to remove the outliers from the sampling frame, review the items separately, and then add the results for the outliers back to any statistical estimates.

A detailed discussion of the many available methods for detecting outliers is beyond the scope of this toolkit. A simple approach to assessing whether outliers may require further review is described below:

1. Pull a draft sample using a tentative design.
2. If the potential outlier is not selected, then replace the largest selected transaction with the potential outlier.
3. Estimate the total paid amount in the frame using the paid amounts in the draft random sample. (This is not a review of the sample. Rather, it is using the paid amounts or some other value that is thought to track potential error amounts.)
4. If the precision of the resulting estimate is reasonable, then the risk associated with the outlier is likely low.
5. Regardless of the result, discard the draft sample.

MISSING DATA: IRRELEVANT VERSUS IRRESOLVABLE SAMPLE ITEMS

Sometimes a sample contains items that are not relevant to the objective of the sample or are relevant but cannot be resolved as correct or in error. The best approach for handling these types of items depends on how the results will be used and the exact reason the sample item cannot be resolved. Given the complexity of the area, this section will provide only a short introduction to the topic.

For the purpose of calculating overpayment totals, irrelevant items are similar to items that are correct (i.e., both represent items that have no recoverable overpayment associated with them). The situation

[7] This discussion assumes that the outlier is a very large positive error. If very large positive and negative errors are possible, then the upper and lower limits of a standard confidence interval may be unreliable.

is more complicated when the goal is to calculate a dollar or count error rate out of relevant items in the frame rather than the frame as a whole. A statistician will be able to calculate these rates but needs to be alerted to the issue.

As an example of this situation, suppose investigators are interested in a potential fraud committed by a specific nurse at a clinic. The investigators pull claims related to that nurse; however, once they start reviewing their sample, they find that they did not account for the days that the nurse was not in the office. The claims handled by the other nurses are not relevant to the investigation of the target nurse. When calculating the total overpayment for the target nurse, the claims associated with all the other nurses can be treated as having $0 in overpayments for the target nurse. This coding reflects the actual liability for the target nurse for these claims.

The solution is not as straightforward when calculating the error rate. Suppose that a statistician estimates that the nurse submitted $50,000 in false claims out of $100,000 in claims paid in the frame. A simple yet incorrect error rate calculation would be 50 percent (i.e., 50,000/100,000). However, this approach would underestimate the error rate, because it would treat claims from other nurses as though they were from the target nurse. A statistician could calculate an adjusted error rate that applies only to the relevant claims. The specific nature of this calculation is beyond the scope of this toolkit.

Unresolvable claims are relevant claims that cannot be identified as correct or incorrect. There are no easy solutions for these types of claims. For example, suppose that all of the claims are for the target nurse but that half of the claims are missing key documentation because of a flood the previous week. Further suppose that the lack of documentation alone is not sufficient to show the claims to be in error. In the example of the irrelevant claims, we were able to calculate an unbiased overpayment estimate by treating the claims from the other nurses as having $0 in overpayment. This approach may be viable for unresolvable claims, but it could substantially understate the amount of fraud present. Conversely, one cannot necessarily assume the claims with lost documentation are the same as claims that can be reviewed. Such an assumption would raise both legal and technical questions. Because there is no easy answer in these cases, careful collaboration with a statistician is critical.

SPARES (Sample Review Step Description)

A spare is an extra sample item that is reviewed in the place of one of the original sample items. Those new to statistics often believe that spares can help increase the estimated error totals calculated from a sample. In fact, the primary use of spares is to improve the precision of the sample in response to the identification of irrelevant or irresolvable sample items. For this limited gain, the cost can be a design that is more difficult to defend, implement, and explain. A statistician may be able to identify alternative methods to handle potential missing or irresolvable items that are technically sound and avoid the difficulties associated with using spares.

STRATIFICATION (Sample Design Step Description)

General Information

Stratification involves the separation of the frame into separate non-overlapping parts and the selection of a fixed number of sample items from each part.[8] A general survey of the topic can be found on pages 180 through 206 in the GAO sampling guide. In a simple example, a provider has three facilities and a team wishes to review transactions at each facility to make one overall

[8] For the stratification to be valid, each item in the frame can appear in only one stratum.

overpayment estimate. The team could split the provider’s transactions into three strata (one stratum for each facility) and select 50 transactions from each stratum.

Stratified designs have several potential advantages:

- By grouping items that are similar, a stratified design may be more precise than a simple random sample with the same sample size. For example, suppose a frame contains both managed care and fee-for-service transactions. The dollar amounts and error rates between these groups may be very different. Stratifying by claim type would keep these differences from harming the precision of the sample.
- If the sample sizes for the individual strata are large enough, then a stratified sample makes it easier to calculate separate estimates for the individual stratum. When planning to calculate an estimate for an individual stratum, the sample size for that stratum should be handled in the same way it would with a simple random sample.
- Stratification allows a team to focus on small areas (e.g., a small, high-risk office location) that may not otherwise be covered by a random sample. A caveat to this approach is that over-representing a small area will tend to result in a less precise estimate.
- It may be easier to explain the results if certain types of claims are kept separate. For example, suppose that only one of the three facilities in the above example has any overpayments. With a simple random sample, one could incorrectly believe that the overpayment findings for the problematic facility are also being applied to the facilities without any overpayments. A similar issue could arise if a frame includes two very different types of durable medical equipment.

Despite these advantages, there are some limitations to stratified designs:

- Designing and implementing a stratified sample requires additional care and expertise.
- An ineffective stratified design may actually be less precise than a simple random sample of the same size.
- On technical grounds, stratified designs are generally just as supportable as simple random samples. However, opposing parties may take advantage of the unintuitive nature of statistics and the increased complexity of stratification to cast doubt on a stratified design; even if these challenges are invalid, they could be seen as persuasive to a jury.

An efficient stratified design can certainly improve the precision of an estimate. The next sections provide general guidance on how to implement an efficient stratified design.

Choice of Stratification Variable

All else being equal, strata should be selected so that the amounts being measured (e.g., overpayments) differ as little as possible within each stratum. Common choices for strata may include paid amount, service code, location, and time.

Choice of the Number of Strata

There are diminishing improvements in precision for each additional stratum added to a sample. As Cochran points out in *Sampling Techniques*, studies have shown minimal gains in precision beyond six strata. Adding more strata will tend to improve the precision but will also increase the complexity of a design and may require more resources.

While it is possible to construct a valid design with as few as two sample items in each stratum, small stratum sample sizes (e.g., fewer than 30) can be difficult to defend for both practical and technical reasons.

Choice of the Strata Boundaries

Basing stratification on a continuous variable (e.g., a dollar value rather than a location) requires selecting strata boundaries. There are several advanced methods for setting strata boundaries. These methods are not included in RAT-STATS but may be available in other statistical software packages. One simple approach is to construct strata that contain an equal number of items. This approach helps avoid designs that are unnecessarily unbalanced but does not in any way guarantee optimal efficiency. A more detailed discussion of this topic is outside the scope of this toolkit.

If strata are defined based on categories (e.g., service codes), then grouping categories together may simplify the design. For example, suppose there are 100 different codes within the sampling frame, but 3 of these codes are considered to be high risk. It would likely be impractical to stratify by code and create 100 different strata. Instead, it may be easier to place the 3 high-risk codes into their own stratum and then group the remaining 97 codes into a small number of combined strata.

Allocating Samples

Once the frame has been broken into separate strata, the next decision is how many sample items to pull from each stratum. This is described in more detail below.

Option 1 — Proportional allocation

The easiest way to allocate sample items across strata is to set the sample size in each stratum so that it is proportional to the number of items within the frame. For example, suppose there is a frame with 1,000 records that is split into 3 strata of 200, 300, and 500 records. For a sample of 200, the proportional allocation would be calculated as (200/1,000) × 200 = 40 items for the first stratum, (300/1,000) × 200 = 60 items for the second stratum, and (500/1,000) × 200 = 100 items for the third stratum. The benefit of this approach is that the design will almost always work as well or better than a simple random sample of the same size. The disadvantage is that this approach may not be as precise as methods that emphasize higher impact strata.

Option 2 — Convenience allocation

In some cases, the sample items will be allocated for reasons other than statistical efficiency. For example, the review may have dual goals of calculating an overall overpayment amount for a provider and providing estimates for three separate offices. If one of the offices is very small, then pulling enough sample items to calculate an estimate for the smaller office may result in a less precise overall estimate. The key point is that allocating the sample for convenience or some other secondary goal may result in a less precise overall estimate than pulling a simple random sample of the same overall size.

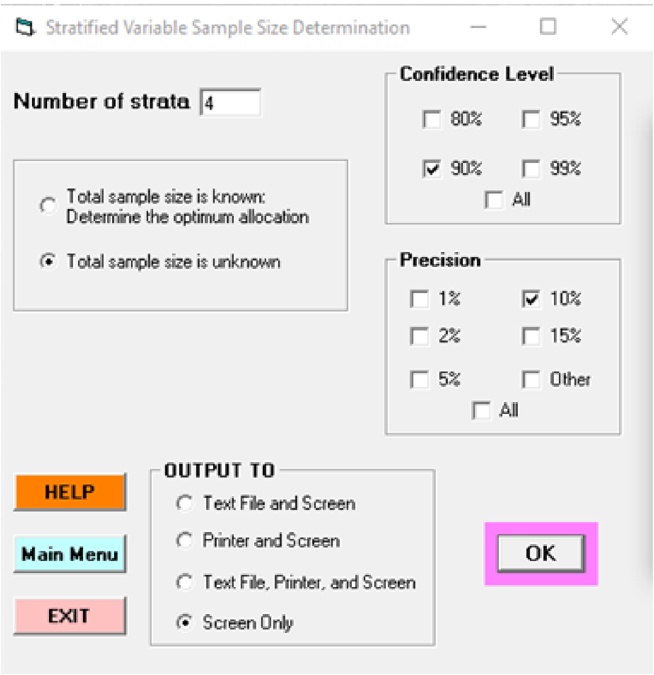
Option 3 — Neyman allocation

A Neyman allocation leverages knowledge about the distribution of error amounts in the sampling frame to improve the efficiency of the sample design. However, this method may perform worse than proportional allocation if one's understanding of the sampling frame is incorrect. This allocation method is included in RAT-STATS and many other common statistical packages.

The explanation below summarizes the RAT-STATS input fields needed to implement the method. Many other statistical packages have similar tools.

Number of strata — Enter the planned number of strata in this field as shown in Figure 3.

Figure 3



Stratified Variable Sample Size Determination

Number of strata 4

○ Total sample size is known: Determine the optimum allocation

● Total sample size is unknown

Confidence Level

□ 80% □ 95%

☑ 90% □ 99%

□ All

Precision

□ 1% ☑ 10%

□ 2% □ 15%

□ 5% □ Other

□ All

OUTPUT TO

○ Text File and Screen

○ Printer and Screen

○ Text File, Printer, and Screen

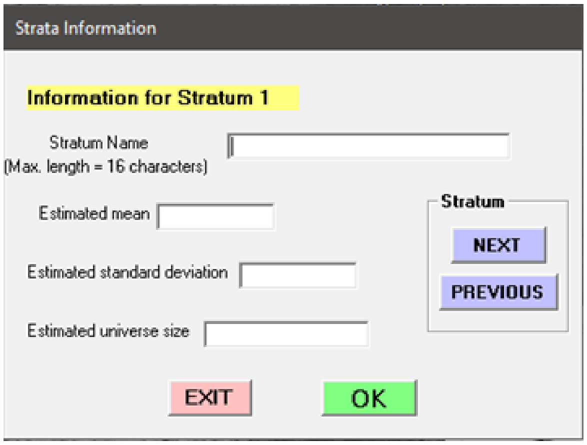
● Screen Only

Confidence level—The confidence level represents the expected percentage of time that the calculated confidence interval will contain the target population quantity. The higher the confidence level, the larger the sample size needed.

Precision—This measure represents the maximum width of a confidence interval one would be willing to accept. For example, if the estimate is 1,000, then a 10-percent precision would arise from a confidence interval ranging from 900 to 1,100. The smaller the precision, the larger the sample size needed.

After entering the necessary information and clicking OK, a new screen will appear that will ask about three quantities for each of the selected strata (the estimated mean, the estimated standard deviation, and the estimated universe size). This screen requires knowledge about the unknown distribution of error amounts. Because this distribution of error amounts is unknown, one must use the available information. One approach is to use the results of a previous sample in a similar area. Another approach is to select and review a small probe sample from the sampling frame. There are other approaches as well, but all approaches require either a reasonably good understanding of the data or significant technical expertise as shown in Figure 4.

Figure 4



Stratum Name—This field is merely a label and can be given any name one desires.

Estimated mean—For whatever quantity being measured (e.g., an overpayment), one needs to provide an approximate value for the average of that quantity for the given stratum.

Estimated standard deviation—The same as above, but now one needs to enter an approximate value for the standard deviation.

Estimated universe size—In practice, this quantity is not an estimate but rather the total number of items within the given stratum.

Appendix H: A Selection of Resources

This section includes resources that may be useful to explore to learn more about statistical sampling. The list is by no means exhaustive. These resources cover the basics of statistical sampling but do not provide information about more complex methods.

- U.S. General Accounting Office, Program Evaluation and Methodology Division. 1992. *Using Statistical Sampling*. GAO/PEMD-10.1.6.

  This is a good general summary of the basic principles behind survey sampling. In addition, it introduces a wide range of more complex methods, such as regression estimation and acceptance sampling. This is one of the more easily understandable resources about sampling available.

- Centers for Medicare & Medicaid Services. 2016. *Medicare Program Integrity Manual*. Pub. No. 100-08, chap. 8.

  This publication outlines the steps required for extrapolations performed by various Medicare program integrity contractors. Although it does not apply to MFCUs, it contains useful tips about designing robust and defensible samples.

- Cochran, William G. 1977. *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons; Kish, Leslie. 1995. *Survey Sampling*, 3rd edition (rev.). New York: John Wiley & Sons.

  Sampling Techniques is one of the seminal textbooks on survey sampling. While additional methods have been developed since the publication of Cochran's classic textbook, the proofs outlined in *Sampling Techniques* are just as valid today as when they were initially written. Kish's *Survey Sampling* is similar in coverage and scope. Originally published in 1965, it has been updated more recently than *Sampling Techniques*. Both textbooks are more technical than the other resources on this list, but they are still invaluable tools for understanding how standard sampling methods work.

- Levy, Paul S., and Stanley Lemeshow. 2008. *Sampling of Populations: Methods and Applications*, 4th edition. New York: John Wiley & Sons; Rao, Poduri S.R.S. 2000. *Sampling Methodologies with Applications.* London and New York: Chapman & Hall/CRC Press; Thompson, Steven K. 2012. *Sampling*, 3rd edition. New York: John Wiley & Sons.

  These references are just three examples of more recent sampling textbooks that provide introductions to sampling methods. *Sampling of Populations* has a greater emphasis on examples than mathematical proofs. *Sampling Methodologies with Applications* contains the most detailed information on nonresponse and nonparametric methods. *Sampling* has the most extensive coverage of adaptive designs and network sampling.