

Applying Effective Sampling Techniques for the Senior Community Service Employment Program.

The data validation process is more than a grant requirement. When executed thoughtfully, the process of validating program data against source documentation can provide unique short-term and long-term benefits. The data validation process can identify areas of improvement in your case management process by identifying fields with high amounts of errors or that lack significant documentation. A consistent focus on improving errors on key data elements can translate into increased program performance on core measures that rely on accurately reported data elements. This post will describe how you can utilize sampling to achieve a data validation process that meets grant requirements while also supporting continuous program improvement. Various sampling methods will be described as well as the overall procedures used by the program in past years to facilitate data validation. We will also highlight some tools to assist with establishing your own sampling methodology.

Designing your Data Validation and Sampling Plan

Developing your data validation process involves the following broad steps.

1. **Define your validation objectives.** In addition to meeting the grant requirement, you can use the data validation process to identify operational improvements that can translate into increased performance. These goals can help you make strategic choices as you define the methodology.
2. **Determine your resource constraints.** The sampling procedures discussed can enable you to generate valuable insights about your participants without the burden of assessing every participant. However, some strategies can be more time-consuming than others. While many base analytic tools, such as Excel, can be used, consider researching other statistical and survey software tools for supporting your sampling methodology.
3. **Determine your sampling methodology.** Determine a process for calculating a sample size. Consider having a minimum and maximum number of people sampled based on your population size. Consider augmenting the random selection process for your sample methodology to allow you to get more operationally relevant information. Document the features of your sampling methodology in a central document.

4. **Review and refine the sampling approach.** Document your experience with the sampling process and the results obtained. Consider conducting an annual review and potential update of the procedure prior to executing it again.

In the next sections, we will take a deeper look at how your sampling process relates to the broader data validation process.

Sampling Techniques and Considerations

Sampling refers to the process of selecting a representative subset of a group of people to estimate characteristics of a larger group. Generating a sample for a population is often more cost effective than validating all records for all participants, even with relatively small population sizes. Randomizing the selection process for this subset ensures that the rate of records with data inconsistencies derived from the subset of sampled case are not negatively influenced by unknown or unobserved factors. DOL encourages grant recipients to incorporate a comprehensive random sampling methodology and procedure for reviewing the source documentation of participant files but does not spell out a specific sampling methodology. There are several factors to consider when outlining a sampling methodology that maximizes the value for your organization.

Sample Size

There are several different statistical techniques available to determine sample size. Determinations of the minimum sample size should depend on the level of precision needed in your estimate and the level of representativeness of the sample. Precision in this context refers to the consistency and reproducibility of a result and is not related to the actual error rate directly. As the population increases, your sample does not have to increase at the same rate to maintain the accuracy of your estimates.

A commonly used version of this formula for calculating sample size (n) is below:

$$n = \frac{\frac{z^2 * p(1 - p)}{\epsilon^2}}{1 - (\frac{z^2 * p(1 - p)}{\epsilon^2 N})}$$

Where:

- N is the population size
- z represents the z-score, a metric used to determine the confidence interval of your measure,
- p represents the proportion of the population holding the relevant characteristic and

- $\mathcal{E}$ represents the margin of error for your estimates.

For a population of 500, we would calculate the sample size to be 218 using a z-score of 1.96, a p value of .5 and an e value of 0.05. There are a variety of resources on the internet to help you calculate sample sizes, such as the following [sample size calculator](#) that allows you to adjust various parameters.

The table below uses the sample size formula employed by the legacy version of SCSEP's case management system (SPARQ) to give you a general sense of sample sizes compared to the population. Note that the legacy application used two different calculations to determine the sample for the eligibility and performance samples.

Number of Participants	Sample Size for Eligible Population	Sample Size for Performance Population ⁱ
50	21	23
100	35	42
200	54	70
500	73	107
1000	115	187
2000	130	230
5000	139	250

Sampling Techniques

While random sampling can make your estimates more generalizable, the small sample sizes may omit key examples that are relevant for your operation. An additional way to improve the operational value of your sample is to divide your population into separate categories prior to selection. **Stratified sampling** uses these different categories, or strata, to ensure that the sample contains relevant subpopulations of participants. Region, race and gender are examples of different strata that can be applied to the sampling frame.

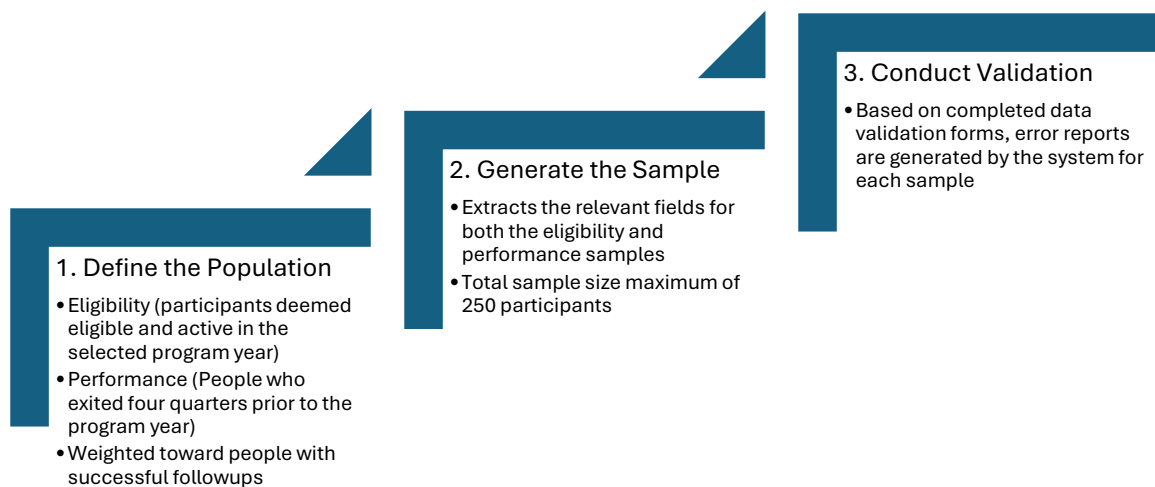
In contrast, there may be other less homogeneous groupings that you could use to develop your sampling frame. **Cluster sampling** involves dividing the population into naturally occurring groups, or clusters, such as geographical regions or organizational units. Organizing your sampling around sub-grantees would be an example of cluster sampling. A random sample from each cluster is then selected, and all members within those clusters are included in the sample. This method is particularly useful when it is more feasible and cost-effective to gather data from clusters rather than individuals.

Most commercial statistical software packages, such as SPSS and STATA, have built-in functionality that allows you to generate random samples. Excel also has specific tools to allow you to randomly select from lists and calculate sample statistics. Online survey

tools, such as Qualtrics and Survey Monkey, also provide functionality to sample participants as well as generate forms needed to collect results of the data validation process.

Example of Data Validation Process from SCSEP Performance and Quarterly Reporting System (SPARQ)

Figure 1: Overview of legacy Sampling Process



The legacy application for SCSEP included a data validation framework that applied many of the concepts discussed above. The description below is meant to illustrate an example of a data validation sampling process and does not have to be mirrored by your data validation procedure.

The process began by generating two stratified random samples based on separate criteria for participants who had gone through eligibility determination and became active, and participants who exited in the fourth quarter before the exit quarter of the program year being sampled for validation. The performance sample was also weighted toward participants who had completed more than one follow-up successfully.

From these populations we determined the sample size by calculating the sample needed to yield a 95% confidence interval in our estimates. For the eligibility sample, the system selected participants at random to populate the sample. The performance sample's participants were selected in multiple rounds that prioritized participants with successful follow-ups to participate in the sample. Both samples had an artificial maximum of 250 participants. This served to ensure the samples were manageable without sacrificing the accuracy of the estimates.

Once the samples were drawn, grantees validated data using the forms, and an error report summarizing the results was created.

Conclusion

Accurate data is the foundation of program excellence. A comprehensive data validation procedure can enhance data quality as well as program operations. As you continue your development of a data validation process, we hope that the information presented helps you maximize the value of this data management process.

Prior versions of data validation in SCSEP have used 0.95 for P and 1.96 for z. Differences in the sample sizes were caused by slight variations in the calculation. The specific sample size formulas are explained in the data validation technical specifications.