

Notes:

Sampling

Key Learning Points

1. Describe the importance of sampling.
2. Explain how to sample data.
3. Utilize sampling in improvement projects.

What is Sampling?

Sampling is the process of collecting a portion or subset of the total data that may be available to find enough valid “information” to make a decision. All of the data is often referred to as a population or a universe. The purpose of sampling is to draw conclusions about the population using the sample. This is known as statistical inference.

Sample Size

Sample size is determined by establishing a confidence level (usually 95%) you would like to have, deciding on the degree of precision wanted in the value that will be inferred (e.g., population average), and an estimate of the variation in the population.

When to Sample

Sampling is used when:

- Collecting all of the data is too costly
- Full data collection may be a destructive process

- Sound conclusions can be made from a relatively small amount of data

Sample Size Can Change

Note what happens to the sample size if you:

- Are willing to be less precise in the value you wish to infer: it goes down
- Are willing to be less confident: it goes down
- Have a wider variance in the population: it goes up

Sampling Equations

The equations used in sample size calculators take these factors into account for both continuous and discrete data.

In sentence format, they look like this: Sample size = $\frac{[(\text{confidence} \times \text{variation}) / \text{precision}]^2}{}$

Considerations for Sampling Situations

Purpose / Aim	Considerations
Describe or Quantify Characteristics	<ul style="list-style-type: none"> ▪ Sampling Scheme <ul style="list-style-type: none"> - Random - Stratified Random ▪ Precision Required ($\pm\Delta$) ▪ Amount of Characteristic's Variation ▪ Confidence Level (95%) ▪ Sample Size

Qualities of a Good Sample

Bias in a sample is the presence or influence of any factor that causes the population or process being sampled to appear different from what it actually is. Bias is introduced when data is collected without regard to key factors that may influence the population or process.

Representative

In a representative sample, the data collected should accurately reflect a population or process. Representative sampling helps avoid biases specific to segments of the proposed process or population under investigation.

Random

In a random sample, data are collected in no predetermined order, and each element has an equal chance of being selected for measurement. Random sampling helps

Notes:

to avoid biases specific to the time and order of data collection, operator, or data collector.

Notes:

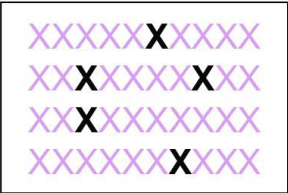

Components of a Sampling Strategy

- Sampling Scheme
 - What is the best way to get a representative sample?
- Measurement Resolution
 - Does the measurement have enough resolution?
- Delta to Detect
 - What change in performance is to be proven?
- Sample Size
 - How many samples do you need to collect?

Random Sampling

The key to random sampling is that each unit in the population has a known equal probability of being selected in the sample. Using random sampling protects against bias being introduced in the sampling process.

In general, random samples are taken by assigning a number to each unit in the population and generating random numbers to pick from the list. Absent knowledge about the factors for stratification for a population, a simple random sample is a useful first step in obtaining samples.

Population	Sample	Description
		Each unit ("X") has a known equal probability of being selected in a sample.

Random Sampling Example

An improvement team in Human Resources wanted an accurate estimate of what proportion of employees had completed an annual employee performance plan.

The team used their database to obtain a list of all associates. Each person on the list was assigned a number. A random number generator generated a list of numbers to be sampled, and an estimate was made from the sample.

Stratified Random Sampling

Stratified random sampling is used when the population has different groups (strata) and you need to ensure that those groups are fairly represented in

the sample. In stratified random sampling, independent samples are drawn from each group, and the size of each sample is proportional to the relative size of the group.

Population		Sample	Description
Strata	Units		
Large	LLLLL	L	<ul style="list-style-type: none"> Randomly sample within a stratified category or group. Sample sizes for each group are proportional to the relative size of the group.
Medium	MMMMM MMMMM MMMMM MMMMM	MMMM	
Small	SSSS SSSS SS	SS	

Stratified Random Sampling Example

The manager of an insurance business wanted to estimate the average cycle time for an application process. He knew there were three types (strata) of applications (large, medium and small). Therefore, he wanted his sample to have the same proportion of large, medium, and small applications as the population.

Bias

Data should be used to answer the questions for which it was collected. The following are some types of biases that may occur and of which you should be aware:

- Exclusion: Some part of the scope of the process being investigated has been left out.
- Interaction: The process of collecting the data itself may affect the process being studied.
- Perception: The attitudes and beliefs of the data collectors can sometimes color what they see and how they record it.
- Operational: Failure to follow the established procedures is the most common operational bias.
- Non-Response: Missing data can bias the result. The fact that they are missing is a clue that they are different from the rest in some way.
- Estimation: The formulas and methods used to calculate statistics from the

Notes:

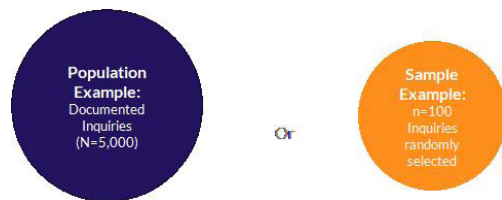
collected data may create certain types of biases.

Sampling Confidence Intervals

In any sampling situation, there is almost always the option of not taking samples, but instead, measuring the entire population. Consider the example below. What is the average resolution time of documented customer inquiries?

One approach to finding the average resolution time is to measure each inquiry. While this might provide the most accurate measure of the average resolution time, for cost and logistical reasons, this might not be practical.

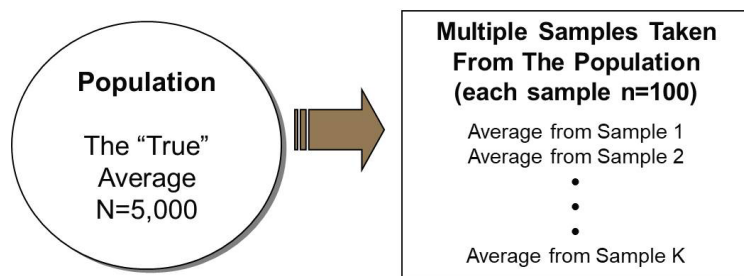
Another more practical approach would be to obtain a representative sample of the 5,000 inquiries, and estimate the average resolution time based on this sample. While this approach costs less, it does not provide as much accuracy. Knowing the degree of accuracy in an estimate is helpful in gauging the usefulness of the estimate in decision-making.



There will be a difference between the “true” average value of a population and an estimated average value from samples.

A confidence interval is a method of expressing the amount of sampling variation in an estimate and includes an interval and a confidence level.

Statistical theory can determine the amount of variation that can be expected in estimates. This is known as a confidence interval. Confidence intervals are stated in terms of an interval or range and a confidence level.



Typical Sample Size

These are rules of thumb to be used to get started. These are minimum sample sizes. Remember, additional data may be worth the additional cost.

Sample size is determined by estimating a confidence level (usually 95%), deciding on the degree of precision wanted in the value that will be inferred (e.g., population or average), and an estimate of the variation in the sample. The equations used in the sample size calculators take these factors into account for both continuous and discrete data.

Notes:

Tool or Statistic	Guidelines Minimum Sample Size
Average	25
Standard Deviation	25
Proportion Defective (P)	100
Histogram, Pareto Chart	40
Box Plot	8
Scatter Diagram	40

Notes:

When Should Sampling Be Used?

Sampling should be used when collecting data to prove root cause when collecting a complete population is too difficult, expensive, or unrealistic. Sampling is usually done during the Analyze step of DMAIC.

Pitfalls to Avoid

Common practices not based on sound statistical principles:

- Fixed Percentage Sampling: Seemingly logical “rules of thumb” such as “always take a ten percent sample” result in under sampling from small populations and over sampling from large populations.
- Judgment Sampling: Instructing a collector to use his or her judgment to select X number of “representative” samples, results in bias.
- Chunk Sampling: Bias and non-representative samples result when you select your sample simply because the items are conveniently grouped—for example, “go to the file cabinet and pull the file of orders where the last name begins with D.” (Note: A technique called “cluster sampling” may sound similar, but it is statistically sound).