**Capstone I**

**Project Report – Assignment 2**

**Sampling Strategies for CLABSI Estimation**

**Fall 2024**

**Team 1**

Jayadurga Machireddy, Lohitha Kolli, Sri Lakshmi Chilekampalli,

BZAN 6360 – Capstone Practicum in Business Analytics I

November 11, 2024

Table of Contents

# Executive Summary

This report looks at the process of choosing the right sampling technique to extract a good sample from a dataset that's very skewed, particularly focused on Central Line-Associated Bloodstream Infections (CLABSI) events. The dataset includes over 5000 observations, with information like patient identifiers, how many times a patient shows up, how many CLABSI events they've had, and whether a CLABSI event occurred or not.

We explored the data to get a sense of how frequently patients and CLABSI events occur, which helped guide us in figuring out how to extract 10% of the data for modeling purposes later on. Out of five probabilistic sampling techniques we reviewed, Simple Random Sampling and Stratified Random Sampling stood out. We picked these two because they seemed to handle the skewness of the data well and offered the best chance to avoid bias.

We also set up workflows to pull samples using both techniques and compared the statistics of those samples to the overall population. This comparison showed how the differences in sampling might affect estimating CLABSI events, and really highlighted how important it is to have a sample that accurately reflects the population.

In the end, the report gives a sampling plan that's aimed at predicting CLABSI occurrences, keeping in mind the unique characteristics of the dataset we're working with.

# Analysis

## Population Profile

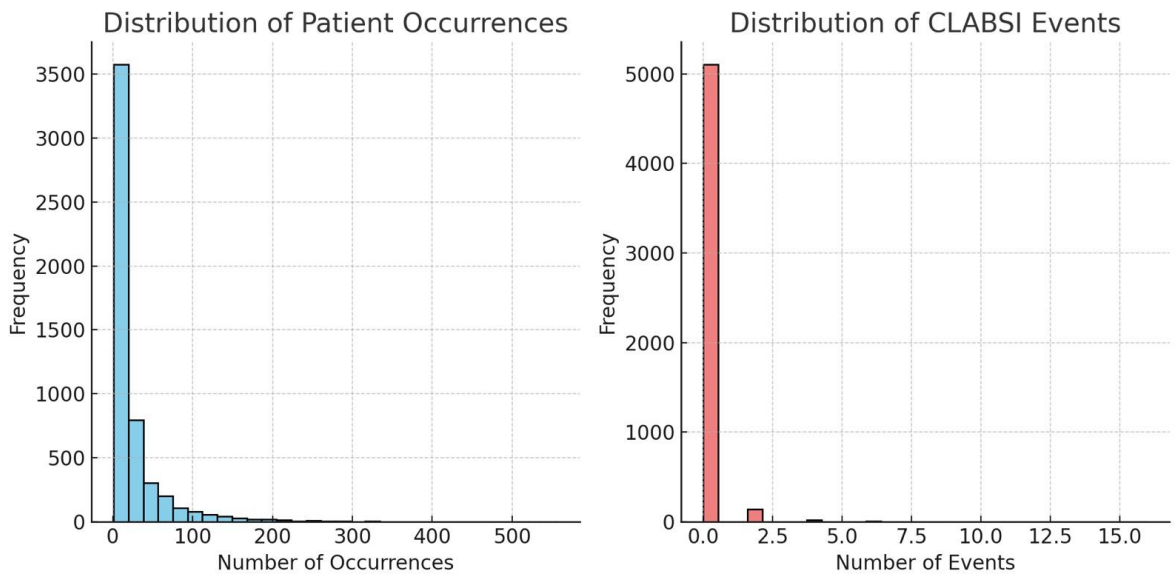
The dataset we're working with has four variables: PatientKey, the number of times a patient shows up in the data (Number of Patient Occurrences), how many CLABSI events a patient has had (Number of CLABSI Events), and whether a CLABSI event occurred (CLABSI Event T/F). In total, there are more than 5000 patients, and around 3% of them have experienced a CLABSI event.

Looking at the data, the distribution is very skewed. Most patients don't appear much, but there are a few that pop up many times, making the dataset right-skewed. It's a similar story with CLABSI events. Most patients don't have any events, but some have a lot, creating a long tail in the distribution.

Here are two graphs that give a better sense of what's going on:

1. Patient Occurrences: You can see from the chart that most patients appear just a few times, but there are some outliers who appear way more often.

2. CLABSI Events: The chart shows a similar pattern—most patients have no or very few CLABSI events, but again, a few patients have many events, which skews the data.




These visuals make it pretty clear that the dataset has a lot of extreme outliers, and that skewness is something we'll need to deal with when choosing the sampling method.

**Distribution of patient occurrence**

| | Mean | Standard Error | Median | Mode | Standard Deviation | Sample Variance | Kurtosis | Skewness | Range | Minimum | Maximum | Sum | Count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 24.507482477742000 | 0.5577308435075680 | 10.0 | 1 | 40.522897719294700 | 1642.1052395684200 | 26.170456913294400 | 4.130196491291564 | 556.0 | 1.0 | 557.0 | 129375 | 5279.0 |

**Distribution of Clasbi events**

| | Mean | Standard Error | Median | Mode | Standard Deviation | Sample Variance | Kurtosis | Skewness | Range | Minimum | Maximum | Sum | Count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.084675127865126000 | 0.007432415495961560 | 0.0 | 0 | 0.540014984747862 0 | 0.291616183752234 0 | 196.941222432182 00 | 10.823637911961700 | 16.0 | 0.0 | 16.0 | 447 | 5279.0 |

## Probabilistic Sampling Techniques

Probabilistic sampling techniques, such as simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and multistage sampling, provide different strategies for selecting samples from a population, depending on the research goal. These methods ensure representativeness and efficient sampling by randomly selecting individuals, organizing populations into strata, selecting clusters, systematically choosing samples at intervals, or combining multiple approaches for more complex scenarios.

The table below offers a brief description of these five techniques, highlighting their advantages and disadvantages:

| Technique | Description | Advantages | Disadvantages |
|---|---|---|---|
| **Simple Random Sampling** | Each individual in the population has an equal chance of being selected. | Easy to implement, unbiased | May miss rare subgroups, especially in skewed data |
| **Stratified Random Sampling** | Divides population into subgroups (strata) and samples proportionally from each. | Ensures representation of subgroups, reduces bias | More complex, requires prior knowledge of strata |
| **Cluster Sampling** | Samples entire clusters or groups, rather than individuals. | Efficient for naturally grouped populations | Can over- or under-represent if clusters vary |
| **Systematic Sampling** | Selects every nth individual after starting from a random point. | Simple to execute, fast | May miss important patterns if there's hidden structure |
| **Multistage Sampling** | Combines methods, often cluster sampling followed by random or stratified sampling within each cluster. | Useful for large, complex populations | Complex to set up, potential sampling bias at different stages |

These techniques offer flexibility depending on the nature of the dataset and the research objectives. For the CLABSI dataset, which has a highly skewed distribution, Stratified Random Sampling is particularly useful to ensure proper representation of the rare CLABSI events, while Simple Random Sampling provides a simpler, unbiased selection, though it may overlook important patterns in the data.

## Case Sampling Technique Selection

When it came to picking a sampling method for our dataset, we looked at five different probabilistic techniques. Each one has its own set of pros and cons, so there were quite a few things to consider. We needed to think about things like how well each method would represent our data, how big the sample size would be, the risk of bias creeping in, and practical stuff like time and cost.

For predicting CLABSI events, we decided to focus on Simple Random Sampling and Stratified Random Sampling. Our dataset is large and pretty skewed, with only a small

percentage of patients experiencing CLABSI, so these two methods seemed to fit best. They do a good job of balancing out the need for accuracy with the need to avoid bias, and both are reliable when it comes to building predictive models.

With Simple Random Sampling, every patient has an equal chance of being picked, so it's straightforward and doesn't require much setup. But because CLABSI events are rare, there's a risk that we wouldn't get enough of those cases in a random sample, which could impact the usefulness of the analysis.

On the other hand, Stratified Random Sampling makes sure both groups—patients with CLABSI and those without—are represented proportionally. That way, we don't miss out on capturing those rare CLABSI cases. The downside here is that we need to know about these groups in advance, so it takes a bit more effort to set up.

Even though each method has its drawbacks—simple random sampling might overlook rare events, and stratified random sampling needs some prep work—these two still gave us the best shot at getting a sample that truly represents our data. And since we already have clear groups in the data (CLABSI vs. non-CLABSI), stratified sampling stood out as a solid choice for making sure our analysis captures what's really going on.

## Simple Random Sampling

Simple random sampling is a technique where every individual in the population has an equal chance of being selected. This method is straightforward and is intended to create an unbiased sample that fairly represents the entire population. In practice, you assign random numbers to each person in the dataset and then pick your sample based on those random values. However, there can be sampling error if the sample doesn't end up reflecting the true diversity of the population.

**Advantages:**

- It's easy to conduct and doesn't require dividing the dataset into subgroups or taking any extra steps. You can go straight to selecting members.
- It's inexpensive and efficient since you're pulling data directly from the population without any need for segmentation.
- This method is unbiased by nature, as each member has an equal chance of being chosen, which usually results in a good overall representation of the population.

**Disadvantages:**

- One drawback is that you need a complete list of the population. If there are any missing members, it can skew the results because only a full list can truly give everyone an equal chance.
- Simple random sampling may not accurately reflect the population if there's a lot of variability. In cases where certain groups are underrepresented, you might miss important insights due to the random nature of the selection.

## Simple Random Sampling Execution

The workflow for executing a simple random sample from a population dataset is illustrated in Figure 6 below.


Our population dataset has a total of 5279 observations with four key columns: **PatientKey**, **Number of Patient Occurrences**, **Number of CLABSI Events**, and **CLABSI EVENT (T/F)**. To pull a 10% sample using simple random sampling, we created a new column and used Excel's RAND() function to generate random numbers for each row. We then sorted the data in ascending order based on these random values and selected the top 10% as our sample.

In **Table 1**, we compare the population and sample characteristics, specifically looking at CLABSI infection incidence. The proportions of CLABSI and non-CLABSI patients in the sample are similar to those in the full dataset, although there's a slight difference in

percentages. This is consistent with the simple random sampling approach, which aims to reflect the larger population's characteristics.

| Category | Population (# of Patients) | Population (%) | Sample (# of Patients) | Sample (%) |
|---|---|---|---|---|
| CLABSI patients | 173 | 3% | 22 | 4% |
| Non-CLABSI patients | 5106 | 97% | 506 | 96% |
| Total Patients | 5279 | 100% | 528 | 100% |

**Table 1 – Simple Random Sampling Population & Sample Profiles**

**Table 2** provides descriptive statistics for both the sample and the full population. As you can see, values like the mean, median, mode, and standard deviation are pretty similar between the sample and the population, showing that the sample is fairly representative. However, the standard error and variance differ a bit, which makes sense given the variability that comes with sampling.

| Descriptive Statistics | # Patient Occurrences (Population) | # Patient Occurrences (Sample) | # CLABSI Events (Population) | # CLABSI Events (Sample) |
|---|---|---|---|---|
| Mean | 24.507 | 25.763 | 0.085 | 0.083 |
| Standard Error | 0.558 | 1.780 | 0.007 | 0.018 |
| Median | 10 | 10 | 0 | 0 |
| Mode | 1 | 1 | 0 | 0 |
| Standard Deviation | 40.523 | 40.892 | 0.540 | 0.419 |
| Sample Variance | 1642.105 | 1672.120 | 0.292 | 0.175 |
| Kurtosis | 26.170 | 14.619 | 196.941 | 28.573 |
| Skewness | 4.130 | 3.384 | 10.823 | 5.210 |

**Table 2 – Simple Random Sampling Population & Sample Descriptive Statistics**

We also ran a hypothesis test to check if there's a statistically significant difference between the mean of the population and the mean of the sample. We used the mean number of patient occurrences for this test. The results, shown in **Table 3**, gave a p-value greater than our alpha level of 0.05. This suggests there's no significant difference between the population mean and sample mean, meaning our sample does a good job representing the population's characteristics.

| | # of Patient Occurrences (Population) | # of Patient Occurrences (Sample) |
|---|---|---|
| **Mean** | 24.507 | 25.763 |
| **Variance** | 1642.105 | 1672.12 |
| **Observations** | 5279 | 528 |
| **Hypothesized Mean Difference** | 0 | |
| **Degrees of Freedom (df)** | 635 | |
| **t Stat** | -0.673 | |
| **P(T<=t) two-tail** | 0.501 | |
| **t Critical two-tail** | 1.964 | |

**Table 3 – t-Test: Two-Sample Assuming Unequal Variances for Number of Patient Occurrences**

# Stratified Random Sampling

Stratified random sampling is a probability sampling method that creates a representative sample by dividing the population into subgroups, or "strata," that are relatively similar. Researchers often use this method when they want each subgroup in the population to be proportionally represented in the sample, allowing for a more accurate understanding of each group's characteristics. For our CLABSI case study, we chose stratified random sampling for a few key reasons:

- The entire population is divided into two main groups—patients who experienced CLABSI and those who did not. This makes it appropriate to structure our sample in the same way, with two strata representing these groups.

- The group of CLABSI patients is quite small, only making up about 3% of the total population. By using stratified random sampling, we can ensure that this small but important group is adequately represented in our sample, allowing us to capture insights that might otherwise be missed.

**Advantages:**

- Samples generated from stratified sampling represent the population more accurately because each subgroup is proportionally represented. This results in samples that are often more reliable and accurate than those produced by other methods.

- Stratified samples aren't affected by the size of the subgroups, so even small groups, like the CLABSI patients in our case, can be adequately captured.

**Disadvantages:**

- Stratified sampling takes more time and requires a clear understanding of the population's structure to divide the data correctly. This need for detailed knowledge and extra steps can make it a more resource-intensive approach.

- There's a risk of overlap or redundancy if subjects are placed in multiple subgroups, which can lead to misleading results.
- This method doesn't work well with heterogeneous data, where groups are hard to define and individuals don't share similar characteristics.

### Stratified Random Sampling Execution

The workflow for executing a stratified random sample from a population dataset is illustrated in Figure 7 below.


*Figure 7 – Stratified Random Sampling Workflow*

The stratified sampling method was executed in Excel, following a process similar to that of simple random sampling. We began by identifying the total number of patients and calculating the appropriate 10% sample size. We then used Excel's RAND() function to shuffle the data randomly within each stratum. This ensured that the sample was randomly selected, while still representing each subgroup proportionally.

The population was divided into two groups based on the **CLABSI EVENT (T/F)** column, where "TRUE" represented patients who experienced CLABSI and "FALSE" represented those who did not. From each group, we extracted a sample based on the population's proportions, as shown in **Table 4**.

| Category | Population (# of Patients) | Population (%) | Sample (# of Patients) | Sample (%) |
|---|---|---|---|---|
| **CLABSI patients** | 173 | 3% | 17 | 3% |
| **Non-CLABSI patients** | 5106 | 97% | 511 | 97% |
| **Total Patients** | 5279 | 100% | 528 | 100% |

**Table 4 – Stratified Random Sampling Population & Sample Profile**

**Table 5** provides descriptive statistics comparing the population and sample in terms of patient occurrences and CLABSI events. The stratified sample generated similar values for mean, median, and mode, aligning closely with the population. Standard deviation and variance were also similar between the sample and population for CLABSI events. However, there were some noticeable differences in kurtosis and skewness, especially for patient occurrences and CLABSI events, likely due to the 10% sample size, which can limit accuracy in capturing extreme values.

| Descriptive Statistics | # Patient Occurrences (Population) | # Patient Occurrences (Sample) | # CLABSI Events (Population) | # CLABSI Events (Sample) |
|---|---|---|---|---|
| Mean | 24.507 | 24.564 | 0.085 | 0.087 |
| Standard Error | 0.558 | 1.965 | 0.007 | 0.023 |
| Median | 10 | 10 | 0 | 0 |
| Mode | 1 | 1 | 0 | 0 |
| Standard Deviation | 40.523 | 45.162 | 0.540 | 0.523 |
| Sample Variance | 1642.105 | 2039.597 | 0.292 | 0.273 |
| Kurtosis | 26.170 | 41.824 | 196.941 | 56.645 |
| Skewness | 4.130 | 5.223 | 10.823 | 7.069 |

**Table 5 – Stratified Random Sampling Population & Sample Descriptive Statistics**

## Hypothesis Testing and Comparison

To further evaluate how well our sample represents the population, we conducted hypothesis testing. The null hypothesis was that the means of the population and sample were similar, while the alternative hypothesis was that they were different.

The results, shown in **Tables 6 and 7**, indicate that the p-values are higher than the standard significance level (alpha = 0.05). This means we failed to reject the null hypothesis, suggesting no statistically significant difference between the population and sample means. These results strengthen the conclusion that the sample is representative of the population.

| | # of Patient Occurrences (Population) | # of Patient Occurrences (Sample) |
|---|---|---|
| Mean | 24.507 | 24.564 |
| Variance | 1642.105 | 2039.597 |
| Observations | 5279 | 528 |
| Hypothesized Mean Difference | 0 | |
| Degrees of Freedom (df) | 615 | |
| t Stat | -0.028 | |
| P(T<=t) two-tail | 0.978 | |
| t Critical two-tail | 1.964 | |

**Table 6 – t-Test: Two-Sample Assuming Unequal Variances for Number of Patient Occurrences**

| | # of CLABSI Events (Population) | # of CLABSI Events (Sample) |
|---|---|---|
| Mean | 0.085 | 0.087 |
| Variance | 0.292 | 0.273 |
| Observations | 5279 | 528 |
| Hypothesized Mean Difference | 0 | |
| Degrees of Freedom (df) | 645 | |
| t Stat | -0.102 | |
| P(T<=t) two-tail | 0.919 | |
| t Critical two-tail | 1.964 | |

**Table 7 – t-Test: Two-Sample Assuming Unequal Variances for Number of CLABSI Events**

## Sampling Technique Comparison

In comparing simple random sampling and stratified random sampling, we observed key differences across various statistical measures. **Tables 8 and 9** highlight these comparisons, showing that while both methods have strengths, stratified random sampling provides mean values closer to the population, capturing the distribution's skewness and kurtosis more effectively.

Analyzing these contrasts helped us draw some key insights:

- **Mean Value Contrasts**: Stratified random sampling produced mean values closer to the population's, making it better for estimating CLABSI rates.
- **Standard Error Contrasts**: Simple random sampling had a lower standard error, indicating more precision in the sample mean as an estimate.
- **Standard Deviation and Variability Contrasts**: Stratified random sampling showed higher standard deviation, capturing a broader range of values but introducing more noise.
- **Kurtosis and Skewness Contrasts**: Higher kurtosis and skewness in stratified samples suggest better capture of extreme values, essential for understanding outliers in healthcare data like CLABSI events.

These observations indicate that while both methods have trade-offs, **stratified random sampling** is well-suited for this dataset, especially when analyzing rare events like CLABSI.

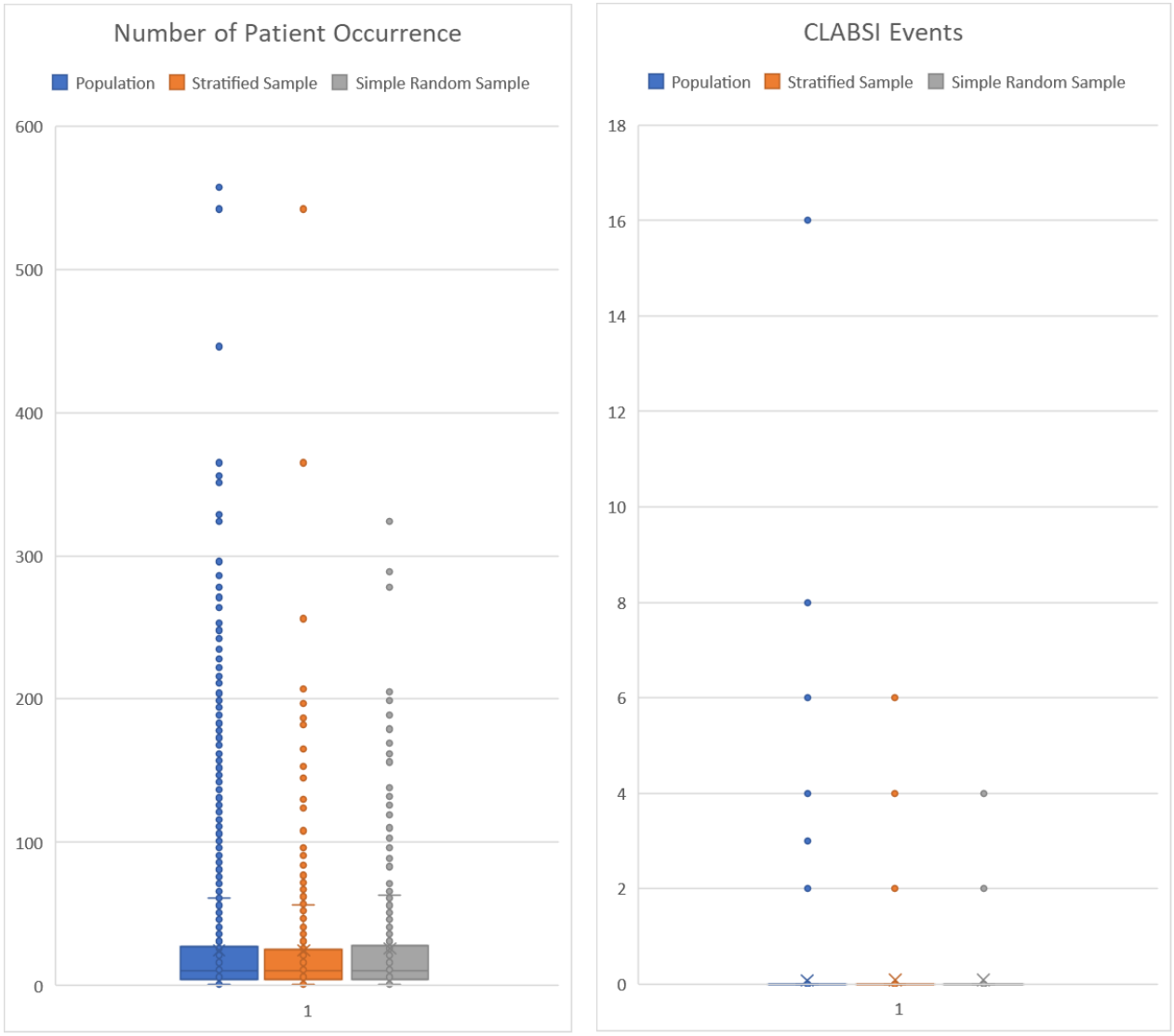



*Figure 8 – Data Distribution of" Number of Patient Occurrence" and" CLABSI Events"*

# Conclusion

Choosing the right sampling technique to get a representative sample from a large, skewed dataset is a critical first step before jumping into predictive modeling. Our analysis showed that several factors need to be weighed when picking a sampling method, each with its own set of strengths and weaknesses. For predicting CLABSI occurrences, both simple random sampling and stratified random sampling emerged as promising options.

Stratified random sampling came out on top when it came to aligning with the population's mean for patient occurrences and CLABSI events. This method was better at capturing the spread of the data, including outliers, and accounting for the asymmetry in our dataset. These qualities are essential when modeling something like CLABSI occurrences, where accurate estimation and representation of rare events make a big difference.

On the other hand, simple random sampling is straightforward, easy to implement, and unbiased, which makes it a good general choice. However, when it comes to capturing the nuances of complex population subgroups, like CLABSI cases, stratified random sampling is better suited. It gives us the accuracy and detail needed for research that requires precise subgroup analysis, making it the preferred choice for this case.