Ramdom Sampling and Performance Analysis

A Formal Report

Cheng Peng

Contents

1 Introduction

Analysis, We carry out an analysis by comparing the performance of three random sampling plans: simple random sampling (SRS), systematic sampling (SS), and stratified sampling based on a large bank load data set as the finite population.

The bank loan data set was provided by the U.S. Small Business Administration (SBA) which contains all historical loans endorsed by SBA from 1987 through 2014. This data set contains 27 variables and 899,164 observations. Each observation represents a loan that was guaranteed to some degree by the SBA. Detailed information about the data set can be found at (https://pengdsci.github.io/datasets/LoanData-description.pdf).

The original data set was split into 9 subsets that are stored on GitHub. We need to load these data sets to R and create a single data set.

The objective of this analysis is to perform an empirical comparison of the three sampling plans using the loan default rate as a reference metric. The North American Industry Classification System (NAICS) code will be used to partition the population into several sub-populations. The discrepancies between the subpopulation default rates and corresponding sample rates under each of the three sampling plans reflect the goodness of the sampling plans.

Since this is an exploratory data analysis, we only use a graph to visually compare the three sampling plans.

In the next few sections, we will review the three sampling plans and perform some data management tasks to define the study population. The three random samples will then be drawn from the study populations.

We will present the comparison using a graphical approach. Some discussions and remarks will be presented at the end of this report.

2 Three Sampling Plans- A Review

In statistical inference, only random samples based on probabilistic sampling can be used to draw statistically valid results. There are several commonly used random sampling plans in practice. In this analysis, we use three of them: simple random sampling (SRS), systematic sampling, and stratified sampling.

2.1 Simple Random Sampling

Simple random sampling is the best sample we can use in statistical analysis. When taking an SRS with size n from a finite population, we assume **all possible** combinations of n data points have an equally likely chance to be selected as the random sample for analysis.

In practice, a sample taken in such a way that each data point in the population has an equally likely chance to be included in the random sample is also called an SRS.

The following figure illustrates the the idea of SRS process.

Image Source: www.questionpro.com/blog/probability-sampling/

2.2 Systematic Sampling

The key step in the **Systematic sampling** is to find the **jump size m** which is approximately equal to the population size (**N**) divided by the size (**n**) of the random sample to be drawn from the population. That is, $m \approx N/n$ (keep only the integral part if it is a decimal). The next crucial step is to take a random number from the first m-th record and then choose every "m-th" record to be a part of the sample. The systematic random sample is valid since its first record is taken randomly.

The following figure shows the idea of a systematic sampling process.

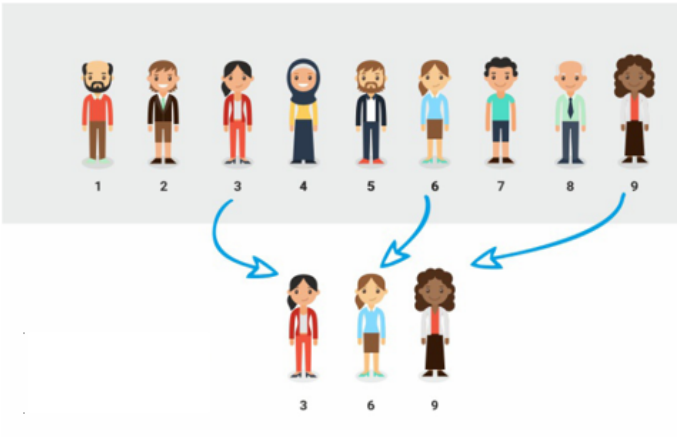


Image Source: www.questionpro.com/blog/probability-sampling/

The above figure has **jump size 3** and the first random subject is the third subject in the population, then the systematic sample is obtained by taking every third subject.

2.3 Stratified Sampling

Stratified random sampling splits the entire population into several subpopulations in some situations in which an SRS is difficult to obtain. For example, when studying a rare disease, it is hard to get an SRS with a sufficient number of diseased subjects in the sample. In this case, we can use the **disease status** as a stratification variable and use its value to split the population into **diseased population** and **disease-free population**. With the two subpopulations, we take two SRS samples from both subpopulations and combine them to obtain a systematic sample.

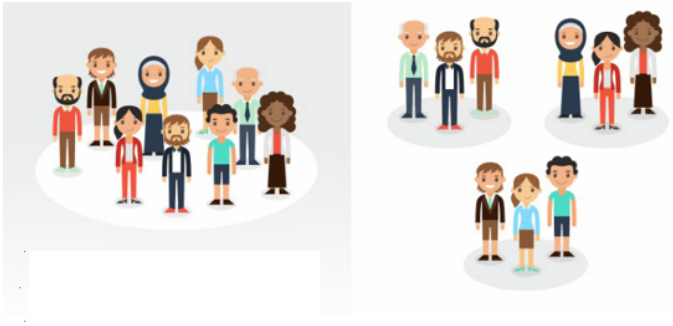


Image Source: www.questionpro.com/blog/probability-sampling/

One important note about stratified sampling is that the subsample must be proportional to the corresponding subpopulation size in order to obtain a combined sample similar to the SRS and systematic samples.

3 Stratification Variable and Study Population

We need to define a stratification variable for stratified sampling. To define a stratification variable or modify an existing categorical variable, we need to make sure each category of the categorical variable should have enough subjects to be sampled. Therefore, we may need to exclude some small categories or combine some categories in a practically meaningful way. The final stratification variable also defines the study population.

3.1 Stratification Variable

There are different ways of defining a stratification variable. For example, we can discretize a numerical variable, use an existing categorical variable, or modify an existing categorical variable. In this analysis, we modify the North American Industry Classification System (NAICS) to define a stratification variable for stratified sampling.

The NAICS is a 6-digit code. We use the first two digits of the code as a basis to define the stratification variable. The population contains 1312 different types of industries according to the 2-digit NAICS code and 1140 of them had less than 900 small businesses.

For the convenience of referring to these tables, I include these two tables in this document.

Description of the first two digits of NAICS.

Sector	Description
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31–33	Manufacturing
42	Wholesale trade
44–45	Retail trade
48–49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services (except public administration)
92	Public administration

Figure 1: List of all industries using the first two digits of the NAICS code

Next, we explore the frequency distribution of the 2-digit NAICS codes and decide the potential combinations of categories with a small size.

0	11	21	22	23	31	32	33
201948	9005	1851	663	66646	11809	17936	38284

42	44	45	48	49	51	52	53	54
48743	84737	42514	20310	2221	11379	9496	13632	68170

55	56	61	62	71	72	81	92
257	32685	6425	55366	14640	67600	72618	229

- 201948 businesses do not have a NAICS code. Since I will use the 2-digit NAICS code to stratify the

Industry default rates (first two digit NAICS codes).

2 digit code	Description	Default rate (%)
21	Mining, quarrying, and oil and gas extraction	8
11	Agriculture, forestry, fishing and hunting	9
55	Management of companies and enterprises	10
62	Health care and social assistance	10
22	Utilities	14
92	Public administration	15
54	Professional, scientific, and technical services	19
42	Wholesale trade	19
31–33	Manufacturing	19, 16, 14
81	Other services (except public administration)	20
71	Arts, entertainment, and recreation	21
72	Accommodation and food services	22
44–45	Retail trade	22, 23
23	Construction	23
56	Administrative/support & waste management/remediation Service	24
61	Educational services	24
51	Information	25
48–49	Transportation and warehousing	27, 23
52	Finance and insurance	28
53	Real estate and rental and leasing	29

Figure 2: List of all industries using the first two digits of the NAICS code and the corresponding loan default rates

population. This variable will be included in the study population that will be defined soon.

- Several categories (21, 22, 49, 55, 92) have relatively small sizes. Since categories 48 and 49 are both transportation and warehouse industries, we will combine the two as indicated in the above two tables.
- As we can see from the above two tables, several industries have different codes. We will combine these codes. In other words, we need to modify the 2-digit code to define the final stratification for stratified sampling.

We now combine the actual 2-digit NAICS codes: **31, 32**, and **33** will be combined and renamed as **313**; **48** and **49** will be combined and renamed as **489**; **44** and **45** will be combined and renamed as **445**. We created a string variable **strNAICS** to represent these modified 2-digit NAICS industries.

3.2 Study Population

Based on the above frequency distribution of the modified 2-digit NAICS codes (the 3-digit codes are combined categories). We use the following inclusion rule to define the **study population**: excluding small-size categories 20, 21, 55, 92, and unclassified businesses with NAICS code 0.

11	23	313	42	445	489	51	52	53	54	56	61	62	71	72	81
9005	66646	68029	48743	127251	22531	11379	9496	13632	68170	32685	6425	55366	14640	67600	72618

The study population has 694216 small businesses across 15 different industries with 29 variables including some derived variables for sampling purposes.

3.3 Loan Default Rates by Industry: Study Population

We now find the loan default rates by industry defined by the stratification variable strNAICS. The loan default status can be defined by the variable MIS_Status.

	no.lab	default	no.default	default.rate
11	10	812	8183	9.0
23	154	15463	51029	23.3
313	126	10438	57465	15.4
42	70	9480	39193	19.5
445	276	28868	98107	22.7
489	123	5939	16469	26.5
51	17	2821	8541	24.8
52	26	2692	6778	28.4
53	44	3904	9684	28.7
54	248	12957	54965	19.1
56	156	7661	24868	23.6
61	24	1552	4849	24.2
62	102	5736	49528	10.4
71	24	3013	11603	20.6
72	89	14882	52629	22.0
81	223	14229	58166	19.7

4 Drawing Random Samples

we are implementing three sampling plans. In each sampling plan, we select 4000 observations in the corresponding samples.

Base R function `sample()` can be used to take SRS samples. To use this R function, we define observation ID numbered from 1 to 694216 so every small business in the study population has a unique ID. The three different samples will be drawn based on these IDs using `sample()`.

For ease of comparison, we keep adding the industry-specific default rates of individual samples to the industry-specific default rates of the study population.

- **Simple Random Sampling**

We simply take random IDs and then identify the records based on the sampled IDs to obtain the SRS sample.

- **Systematic sampling**

The **jump size** is calculated by $m = 694216/4000 = 173.55$. The actual jump size is 173. We use `sample()`random take a record from the first 173 records and then select every 173rd record to include in the systematic sample.

- **Stratified Sampling**

We take an SRS from each stratum. The sample size should be approximately proportional to the size of the corresponding stratum. First, we calculate the SRS size for each stratum and then take the SRS from the corresponding stratum. Then take SRS samples from the corresponding subpopulations.

5 Performance Analysis of Random Samples

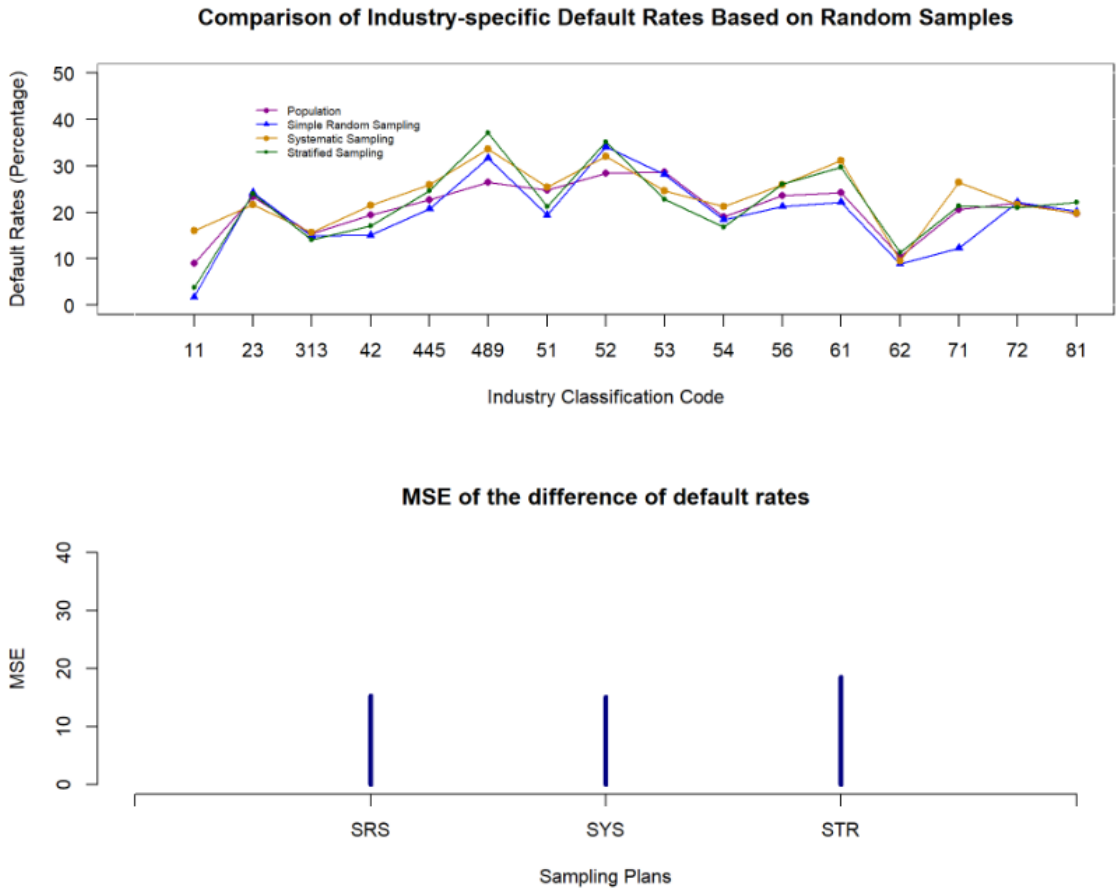
In this section, we perform a comparative analysis of the three random samples. One metric we can use is the default rate in each industry defined by the first two digits of NAICS classification code. That was also used as the stratification variable in the stratified sampling plan.

We have calculated the default rate across the industries in the previous section. That table includes the category with no NAICS classification code. We will use these population-level industry-specific rates as a reference and compare them with the sample-level industry-specific default rates. The summarized table with population and sample level default rates are given below.

	default.rate.pop	default.rate.srs	default.rate.sys	default.rate.str
Agri-forest-fish-hunt (11)	9.0	1.8	16.0	3.8
Construction (23)	23.3	24.2	21.6	24.0
Manufacturing (313)	15.4	14.9	15.6	14.1
Wholesale-trade (42)	19.5	15.1	21.5	17.1
Retail-trade (445)	22.7	20.8	26.0	24.6
Transport-warehousing (489)	26.5	31.7	33.6	37.2
Information (51)	24.8	19.4	25.4	21.2
Finance-insurance (52)	28.4	34.1	32.1	35.2
Real-estate-rental (53)	28.7	28.2	24.7	22.8
Prof-sci-tech-serv (54)	19.1	18.4	21.2	16.9
Admin-support-waste-mgnt-remed (56)	23.6	21.3	26.0	26.1
Edu-serv (61)	24.2	22.2	31.2	29.7
Healthcare-social-assist (62)	10.4	8.9	9.5	11.4
Arts-entertain-rec (71)	20.6	12.3	26.5	21.4
Accommodation-food-serv (72)	22.0	22.2	21.6	21.0
Other-serv(no-public-admin (81))	19.7	20.2	19.8	22.2

First of all, we note that the above table of default rates based on random samples are random. The follow observations are solely based on this random table.

- The sample default rate in some industries have a very large variations comparing with the true default rates at the population level. For example in categories of **agri-forest-fish-hunt**, **Transport-warehousing**, **Edu-serv**, and **Arts-entertain-serv**.
- The sample default rates are close to that of the population rates. We will not test the significance of the differences between the default rates between the population and samples.



The above patterns of industry-specific default rates are also reflected in the following line plot (top panel).

To see the overall performance among the three sampling plans based on these single-step samples (under each of the three sampling plans), we look at the mean square errors of the differences in the default rates between the population and each of the three random samples. The result is summarized in the bottom panel of the above figure. It turns out that the SRS and systematic sampling plans outperform the stratified sampling plan.

A cautionary note: The above-observed pattern about the discrepancy of population and sample rates could be changed significantly across the samples.

6 Discussions and Concluding Remarks

We have implemented the three well-known sampling plans that are commonly used in practice based on a large SBA bank loan data. The NAICS industry code was used to define the study population and the stratification variable for stratified sampling. The difference between population-level industry-specific default rates and sample-level rates was used to compare the performance of the sampling plans.

The comparison results were based on the one-step sample, there could be significant variations. A more reliable approach to obtaining a stable overall performance of the three sampling plans is to take multiple samples and compare the mean MSEs.