Random Sampling Plans and Comparisons
[A Brief Report[]

Cheng Peng

Contents

1 Introduction

This will briefly introduce the steps for taking random samples from a finite population. There are two major types of sampling plans: probabilistic and non-probabilistic sampling plans. The probabilistic sampling plans generate statistically valid data based on which one can perform inferential statistical analysis.

In this project, we consider three probabilistic sampling plans: simple random sampling (SRS), stratified sampling, and systematic sampling. We will use the loan default status as a reference variable to assess the performance of the three sampling plans.

The **Bank loan** data set will be treated as a finite population. After we load the data to R, we will perform some data manipulation and exploratory data analysis to identify a stratification variable for stratified sampling.

2 Identifying A Stratification Variable

We use the 2-digit NAICS code as the stratification variable to define the sub-population.

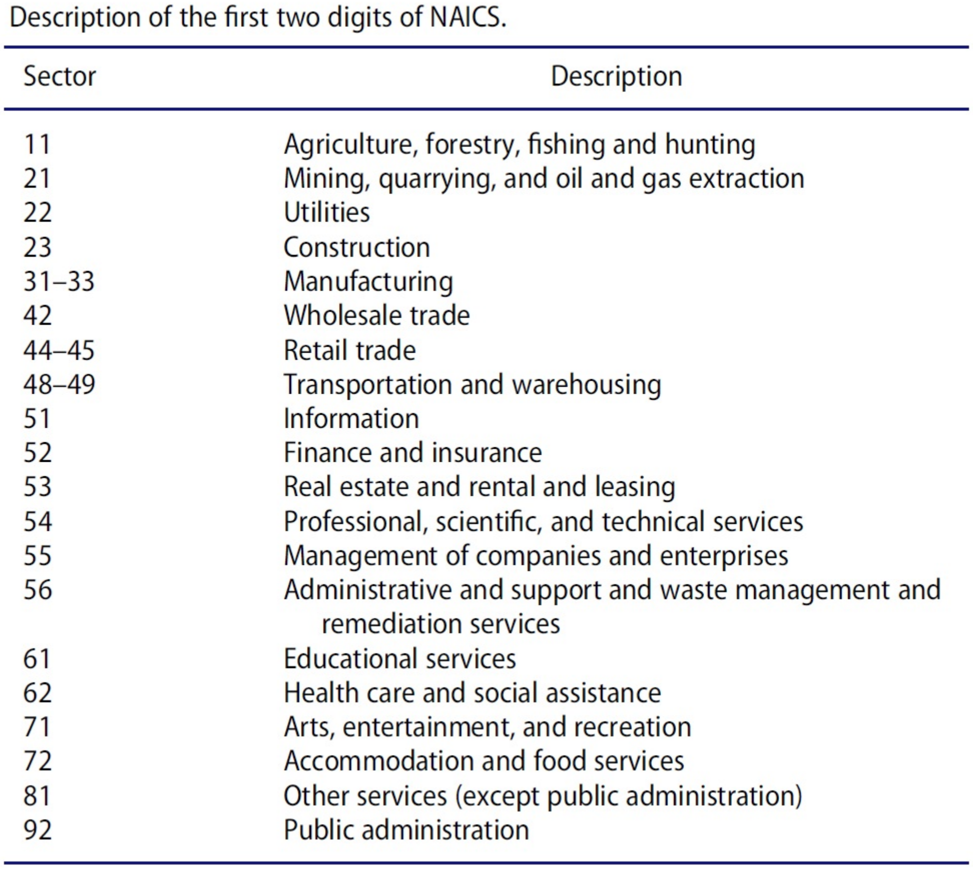
2.1 2-Digit NAICS Codes

We will still use NAICS to create a categorical variable that is operational in practice. Recall that in the last note, we summarized the distribution of NAICS in the following table.

Population.size	Number.of.Industries	Sub.Pop.less.1000
899164	1312	1140

The course web page provides two useful tables. One of the tables https://sta490.s3.amazonaws.com/w03-NAICS-Categories.jpg in the article listed categories based on the first digits of NAICS code. The other related table gives the loan default rate in the corresponding industries https://sta490.s3.amazonaws.com/w03-NAICS-Default-Rates.jpg. You can download these two tables to your local drive and include them in your R Markdown document if you want to practice and reproduce this report.

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.

NAICS.2.digits	Freq
0	201948
11	9005

NAICS.2.digits	Freq
21	1851
22	663
23	66646
31	11809
32	17936
33	38284
42	48743
44	84737
45	42514
48	20310
49	2221
51	11379
52	9496
53	13632
54	68170
55	257
56	32685
61	6425
62	55366
71	14640
72	67600
81	72618
92	229

Several patterns you observe from the above table:

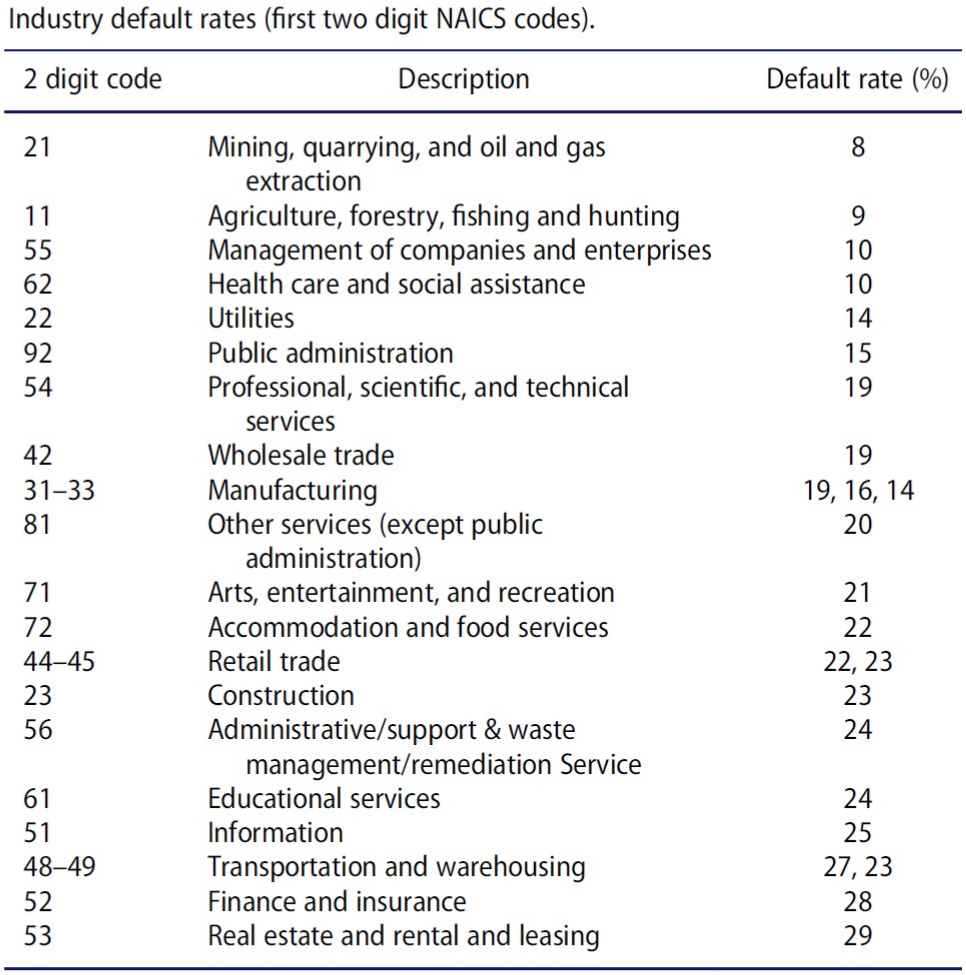
- 201948 businesses do not have a NAICS code. Since I will use the 2-digit NAICS code to stratify the population. This variable will be included in the study population that will be defined soon.
- Several categories (21, 22, 49, 55, 92) have relatively small sizes. Since categories 48 and 49 are both 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.

2.2 Combining Categories

We now combine categories suggested in the above NAICS tables in a meaningful way. Before we combine the NAICS codes, we present an example to illustrate how to combine categories using R.

```
cate.vec0=c(1,4,3,6,7,3,6,5,4,6,4,5,8,9,4,3,4,7,3)  # vector of category labels
cate.vec=c(1,4,3,6,7,3,6,5,4,6,4,5,8,9,4,3,4,7,3)   # a copy of the vector of category labels
labs.2.collapse = c(1,6,7)                          # define a vector to store categories {1,6,7}
logic.vec=cate.vec %in% labs.2.collapse             # TRUE/FALSE ==> match not no-match
cate.vec[logic.vec] = 99                            # if matches (i.e., 1, 5, 7), the value
                                                    # will be replaced by 99
matx=rbind(cate.vec0=cate.vec0, cate.vec=cate.vec)  # check the results
colnames(matx) = 1:length(cate.vec)                 # next kable() function requires a column names
kable(matx)
```

We now combine the following sets of 2-digit NAICS codes: (31, 32, 33) will be relabeled as **313**, (48, 49) as **489**, and (44, 45) as **445**. We use strNACIS to denote the resulting stratification variable.



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

Var1	Freq
0	201948
11	9005
21	1851
22	663
23	66646
313	68029
42	48743
445	127251
489	22531
51	11379
52	9496
53	13632
54	68170
55	257
56	32685
61	6425
62	55366

Var1	Freq
71	14640
72	67600
81	72618
92	229

2.3 Loan Default Rates by Industry

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. The following table gives the default rates by industry.

	no.lab	default	no.default	default.rate
0	281	16799	184868	8.3
11	10	812	8183	9.0
21	0	157	1694	8.5
22	1	94	568	14.2
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
55	1	26	230	10.2
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
92	2	35	192	15.4

2.4 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

So we have defined our study population! The new population size is 694216.

3 Sampling Plans

In this section, we are implementing three sampling plans. In each sampling plan, we select 3000 observations in the corresponding samples. The sample size is slightly less than 5% of the study population size. We use without-replacement sampling plans for all three probabilistic sample plans.

3.1 Simple Random Sampling

We define a sampling list and add it to the study population.

```
## [1] 3000   30
```

3.2 Systematic sampling

We can choose a random integer that is less than the jump size (231 in our case) to guarantee to obtain enough samples.

```
## [1] 3005   30
```

Because the jump size involves rounding error and the population is large, the actual systematic sample size is slightly different from the target size. In this report, I used the integral part of the actual jump size. The actual systematic sampling size is slightly bigger than the target size. We can take away some records from the systematic sample to make the size to be equal to the target size.

3.3 Stratified Sampling

We take an SRS from each individual 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. The following table shows the sub-sample sizes.

11	23	313	42	445	489	51	52	53	54	56	61	62	71	72	81
39	288	294	211	550	97	49	41	59	295	141	28	239	63	292	314

When coding, we use a loop to take simple random samples from each stratum and then combine them to get the stratified sample.

4 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.

4.1 Population-level Default Rates

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 this population-level industry-specific rate as a reference and compare it with the sample-level industry-specific default rates.

Table 7: Population size, default counts, and population default rates

	no.lab	default	no.default	default.rate
0	281	16799	184868	8.3
11	10	812	8183	9.0
21	0	157	1694	8.5
22	1	94	568	14.2
23	154	15463	51029	23.3
313	126	10438	57465	15.4

	no.lab	default	no.default	default.rate
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
55	1	26	230	10.2
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
92	2	35	192	15.4

4.2 Industry-Specific Default Rates based on SRS

For comparison, we construct the following table that includes the industry-specific default rates.

Table 8: Comparison of industry-specific default rates between population and the SRS.

	default.rate.pop	default.rate.srs
Agri-forest-fish-hunt	9.0	10.6
Construction	23.3	24.7
Manufacturing	15.4	11.3
Wholesale-trade	19.5	21.5
Retail-trade	22.7	22.5
Transport-warehousing	26.5	24.8
Information	24.8	19.6
Finance-insurance	28.4	36.2
Real-estate-rental	28.7	34.0
Prof-sci-tech-ser	19.1	20.9
Admin-support-waste-mgnt-remed	23.6	22.9
Edu-serv	24.2	20.8
Healthcare-social-assist	10.4	13.1
Arts-entertain-rec	20.6	18.5
Accommodation-food-ser	22.0	23.8
Other-ser(no-public-admin)	19.7	16.4

Some of the industry-specific default rates seem to be significantly different. More visual comparisons will be given in the next section.

4.3 Industry-specific Rates- Systematics Sample

We will use the sample stratification variable to find the industry-specific rates based on the systematic sample. The following table will include rates of population, SRS, and systematic random samples.

Table 9: Comparison of industry-specific default rates between population, SRS, and Systematic Sample.

	default.rate.pop	default.rate.srs	default.rate.sys
11	9.0	10.6	11.4
23	23.3	24.7	19.4
313	15.4	11.3	18.6
42	19.5	21.5	22.6
445	22.7	22.5	26.2
489	26.5	24.8	27.0
51	24.8	19.6	13.6
52	28.4	36.2	30.6
53	28.7	34.0	22.0
54	19.1	20.9	18.8
56	23.6	22.9	27.1
61	24.2	20.8	16.0
62	10.4	13.1	9.8
71	20.6	18.5	18.6
72	22.0	23.8	20.3
81	19.7	16.4	20.4

It seems that the systematic sample performs better than the SRS sample.

4.4 Industry-specific Default Rates- Stratified Sample

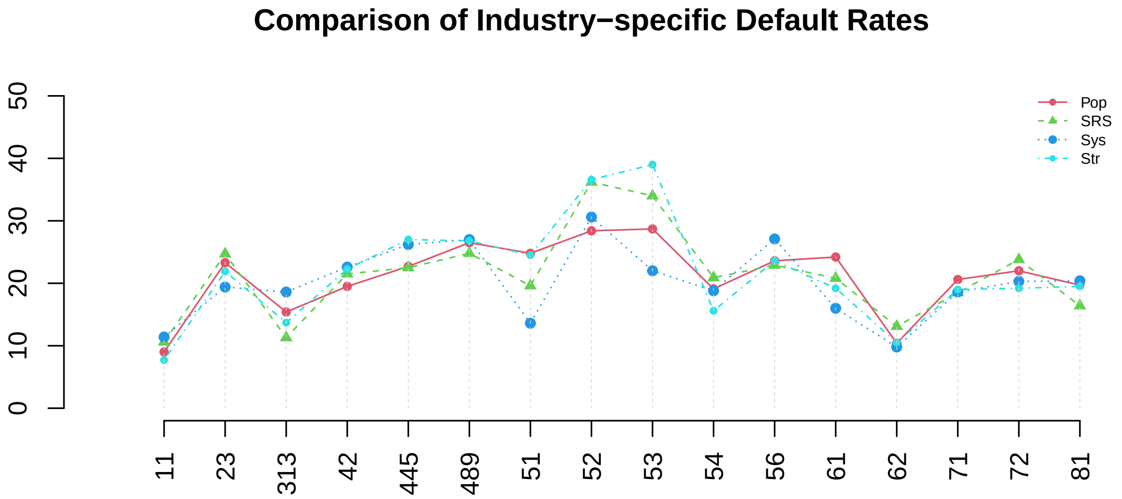
In this section, we put all the information in the following table.

Table 10: Comparison of industry-specific default rates between population, SRS, Systematic Sample, and Stratified Samples.

	default.rate.pop	default.rate.srs	default.rate.sys	default.rate.str
Agri-forest-fish-hunt	9.0	10.6	11.4	7.7
Construction	23.3	24.7	19.4	21.9
Manufacturing	15.4	11.3	18.6	13.7
Wholesale-trade	19.5	21.5	22.6	22.3
Retail-trade	22.7	22.5	26.2	27.0
Transport-warehousing	26.5	24.8	27.0	26.8
Information	24.8	19.6	13.6	24.5
Finance-insurance	28.4	36.2	30.6	36.6
Real-estate-rental	28.7	34.0	22.0	39.0
Prof-sci-tech-ser	19.1	20.9	18.8	15.6
Admin-support-waste-mgnt-remed	23.6	22.9	27.1	23.6
Edu-serv	24.2	20.8	16.0	19.2
Healthcare-social-assist	10.4	13.1	9.8	10.5
Arts-entertain-rec	20.6	18.5	18.6	19.0
Accommodation-food-ser	22.0	23.8	20.3	19.2
Other-ser(no-public-admin)	19.7	16.4	20.4	19.5

5 Visualization - Visual Comparison

In the previous section, we calculated the industry-specific default rates for population, SRS, systematic, and stratified samples. We now create a statistical graphic to compare the default rates among the samples.



For ease of interpretation, I make the following table to map the NAICS codes and the corresponding industries.

Table 11: Industry NAICS codes and the corresponding names

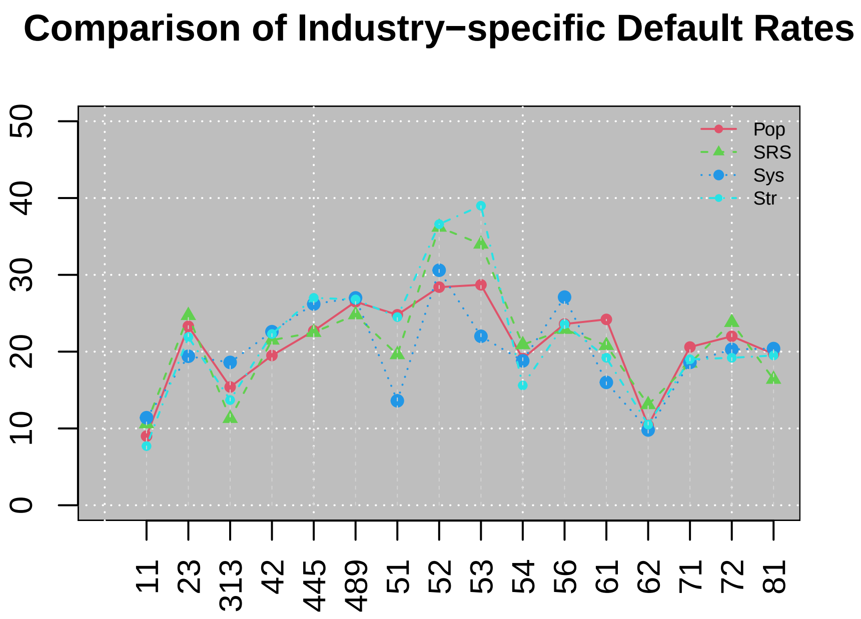
industry.name	industry.code
Agri-forest-fish-hunt	11
Construction	23
Manufacturing	313
Wholesale-trade	42
Retail-trade	445
Transport-warehousing	489
Information	51
Finance-insurance	52
Real-estate-rental	53
Prof-sci-tech-ser	54
Admin-support-waste-mgnt-remed	56
Edu-serv	61
Healthcare-social-assist	62
Arts-entertain-rec	71
Accommodation-food-ser	72
Other-ser(no-public-admin)	81

Since the above graph is based on one-time samples, the variation is not included. Therefore, we need more information to do a meaningful comparison. For example, we can package the code in this Markdown document and take, say 1000, samples based on each sampling plan. We then can find the mean industry-specific default rates and the corresponding variation.

5.1 Critiques of Visual Representation

Visual representation is a key component in effective storytelling. As an example, we critique the figure of performance comparison of the three sampling plans in the previous section and seek improvements for effective representation.

The following figure is modified based on the comparison line plot given in the previous section.

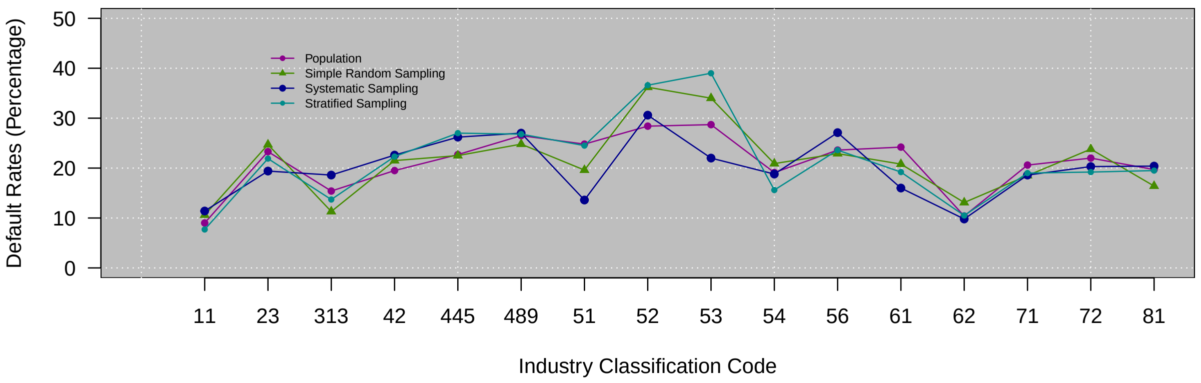


There's a lot of ink here that doesn't convey information relevant to the main point we're trying to make.

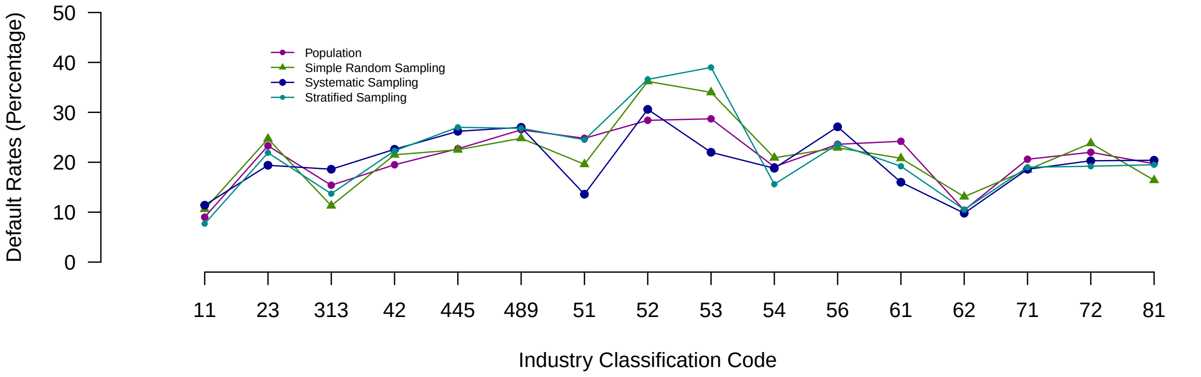
- **Grey background**: not only does it provide absolutely no information. it's also unsightly. After we remove it, we will likely have to darken some of the lines.
- **Grid lines**: it's very unlikely that our audience cares about the exact values at each data point - it's the pattern that matters. The grid lines compete with the pattern we're trying to show.
- **Legend**: It takes time to guess the abbreviation. It also confuses readers.
- **Tick marks**: Why make the reader tilt his or her head to read?
- **Labels of X- and Y- Axis**: labels are missing.
- **Title**: The title should be more specific.
- **Line types**: Use color and line type to differentiate - this will help people who have a color-impaired vision, and also any grey-scale copies of the poster you make (as for handouts).
- **Color coding**: Avoid using use green and red colors in the same graphic to represent different objects. More than 99% of all color-blind people are suffering from a red-green color vision deficiency.

It's easy to make a graph that looks cleaner and has a higher ratio of information-to-total ink:

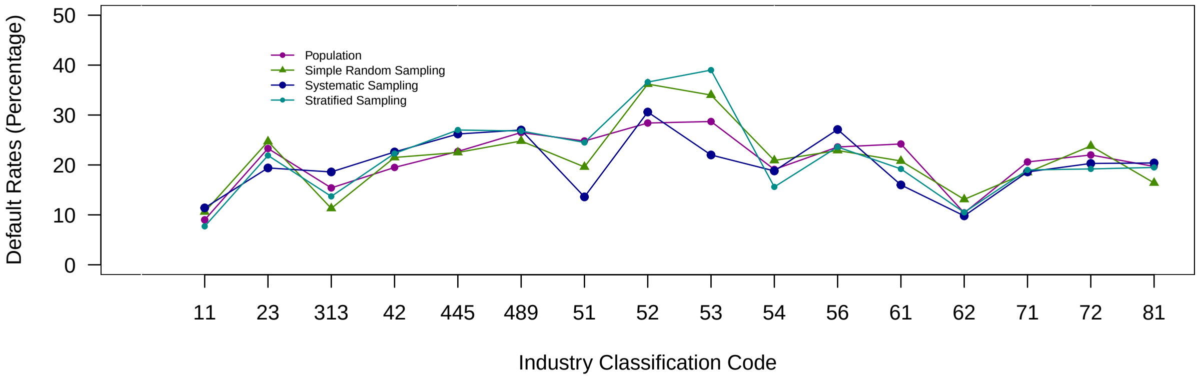
Comparison of Industry−specific Default Rates Based on Random Samples



Comparison of Industry−specific Default Rates Based on Random Samples



Comparison of Industry−specific Default Rates Based on Random Samples


Default Rates (Percentage)
Population
Simple Random Sampling
Systematic Sampling
Stratified Sampling
Industry Classification Code