# MSDS6330 – Statistical Sampling

## HW1 Solutions by Hari Narayan Sanadhya

**Exercise 3.1** - Take a list and draw a simple random sample and a systematic sample from it.

**Solution: -**

Used the dataset provided in the rectangle exercise for this questions solution.

**Simple Random Sample**

We have 100 rectangular fields with ID 1 to 100. The dataset contains the dimensions of the fields along with the area. The problem statement is to estimate mean area of all the 100 rectangles by using a sample of 10 fields. Using the formula " **=RANDBETWEEN(1, 100)**" in excel, got 10 random rectangles and computed the estimated total area using SRS.

|  |  |  |  |
| --- | --- | --- | --- |
| **RandomRectangles** | **Area** | **Sample Mean Area** | **Estimated Total area** |
| 96 | 18 | 8.5 | 850 |
| 89 | 12 |  |  |
| 9 | 1 |  |  |
| 17 | 9 |  |  |
| 71 | 4 |  |  |
| 70 | 18 |  |  |
| 97 | 4 |  |  |
| 68 | 12 |  |  |
| 9 | 1 |  |  |
| 95 | 6 |  |  |

**Systematic Sample**

The variable of interest for the rectangle exercise was Area. I have used Systematic sampling with implicit stratification because implicit stratification ensures that the sample selection will spread across different types of population members and will also help in breaking any periodicity if exists in the dataset.

Step 1- Sorted the data in the ascending order of variable of interest (Area).

Step 2 – Divide the population size (100 in this case) by the desired sample size (10 in this case) to get the sampling interval i.e. i. In our case, i = 100/10 = 10

Step 3 – Draw a random number between 1 and i (i.e. 10). Call it 's'.

Step 4 – Draw the entire sample which would be rectangles with id s, s+i, s+ 2i, s+3i, etc.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample Rectangles** | **Area** | **Sample Mean Area** | **Estimated Total area** |
| 5 | 1 | 6.7 | 670 |
| 15 | 1 |  |  |
| 25 | 3 |  |  |
| 35 | 4 |  |  |
| 45 | 5 |  |  |
| 55 | 6 |  |  |
| 65 | 9 |  |  |
| 75 | 10 |  |  |
| 85 | 12 |  |  |
| 95 | 16 |  |  |

The attached excel sheet contains the implementation of the solution. It has two tabs – SRS and Systematic Sample that shows how was the above solution achieved using excel.



**Exercise 3.2** - In a survey on parks and recreation, 40% of respondents are under 55 years of age, and 60% are 55 and over. The relevant population is known to have 80% under 55 and 20% over. The unweighted survey results indicate that 26% of area residents want more children’s playgrounds in city parks: 50% for respondents under 55 years of age and 10% for respondents 55 and over. If these results are weighted to correct for age, what percentage of area residents want more children’s playgrounds?

**Solution:**

Let's suppose the sample size is 1000 which means based on the sample distribution given, 400 are under 55 years of age, and 600 are 55 and over.

26% of sample (area residents) want more children’s playgrounds in city parks i.e. 26% of 1000 = 260 persons want to have more children's playground.

50% for respondents under 55 years of age and 10% for respondents 55 and over which means 50% of 400 i.e. 200 respondents under 55 years of age wants more Children's playgrounds and 10% of 600 i.e. 60 respondents of 55 and over age wants more Children's playgrounds, the total of which is 260 which matches with 26% of the total population.

Now when we try to generalize this to the population, we have 80:20 ratio of respondents under 55 and over 55. In sample, this ratio is 40:60.

The weight that should be applied for the sample result in case of respondents under 55=80/40 = 2

Similarly, the weight that should be applied for the sample result in case of respondents over 55=20/60 = 1/3

As per the survey results, we found that 200 respondents under 55 want more children’s playgrounds in city parks. If the sample had the same distribution as that of the population, the number of responders under 55 supporting more children playgrounds would have been 2\*200 (weight \* sample result) = 400

Similarly, as per the survey results, we found that 60 respondents over 55 want more children’s playgrounds in city parks. If the sample had the same distribution as that of the population, the number of responders over 55 supporting more children playgrounds would have been 1/3\*60 (weight \* sample result) = 20.

This mean that of the sample size of 1000, 400 + 20 = 420 would have indicated for more children’s playgrounds in city parks. **This means that instead of 26% of the area residents, 42% of the area residents would have shown requirement for more children’s playgrounds in city parks.**