**MSDS 6379**

**Lab 1: Sampling distribution and simple random sampling**

**Submission by Hari Narayan Sanadhya**

**Objectives:**

* Introduce the student to the idea of a probability sample and simple random sample.
* Give the student a preview of the idea of the sampling distribution via a demonstration

**Exercise 1**: Using student responses from the Rectangle Exercise in your asynchronous session, calculate the mean, minimum, and maximum of the sample estimates from all the class students for the judgement samples and SRSs. Compare the mean of all the sample estimates to the population mean which is 738. How many of the student estimates are too high and how many are too low? Enter your responses in the table below:

**Solution**

There are 2 students (student ID – 1 and 3) whose judgement sample average is very high as compared to the population mean of 738.

Sample data provided

|  |  |  |
| --- | --- | --- |
| **Student** | **Judgement Sample Average** | **SRS Sample Average** |
| 1 | 1280 | 780 |
| 2 | 780 | 810 |
| 3 | 1000 | 900 |
| 4 | 690 | 700 |
| 5 | 730 | 660 |

|  |  |  |
| --- | --- | --- |
| **Item** | **Judgement Samples** | **SRS** |
| Mean total area | 896 | 770 |
| Minimum total area | 690 | 660 |
| Maximum total area | 1280 | 900 |
| Number < population mean | 2 | 2 |
| Number >= population mean | 3 | 3 |

For the exercises below, use the table below to craft you responses. This table contains the responses of students to the Rectangle Exercise in a previous section of the Sampling class.

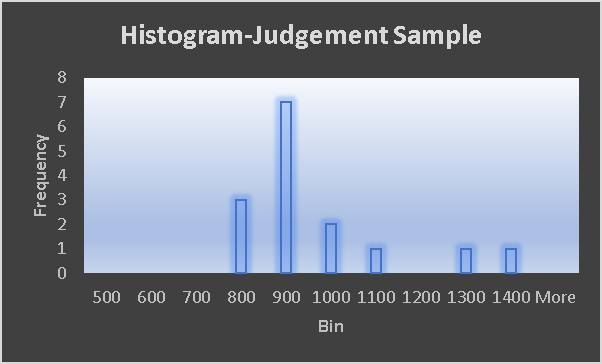
|  |  |  |
| --- | --- | --- |
| **Sample #** | **Judgement Sample** | **Simple Random Sample** |
| 1 | 920 | 750 |
| 2 | 940 | 780 |
| 3 | 850 | 590 |
| 4 | 790 | 780 |
| 5 | 890 | 660 |
| 6 | 780 | 870 |
| 7 | 900 | 830 |
| 8 | 890 | 980 |
| 9 | 1270 | 720 |
| 10 | 810 | 650 |
| 11 | 1040 | 620 |
| 12 | 790 | 700 |
| 13 | 890 | 770 |
| 14 | 850 | 850 |
| 15 | 1320 | 780 |

**Exercise 2**: Calculate the overall mean, maximum, and minimum using each of the total area estimates for the Judgment Samples and SRSs. Enter you responses in the table below.

|  |  |  |
| --- | --- | --- |
| **Item** | **Judgement Samples** | **SRS** |
| Mean total area | 928.6667 | 755.3333 |
| Minimum total area | 780 | 590 |
| Maximum total area | 1320 | 980 |

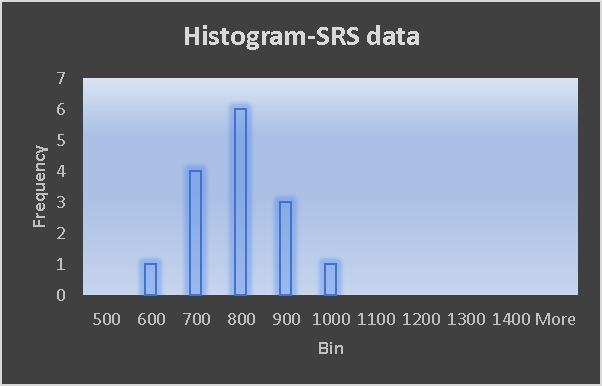
**Exercise 3:** Sketch or use software and create a histogram of the 15 judgement sample values of total area:

Used Excel



**Exercise 4:** Sketch or use software and create a histogram of the 15 SRS values of total area:

Using Excel

****

**Exercise 5:** The mean of estimates of the total area of SRSs is designed to be an estimate of the population total area, i.e., the sum of the areas of all 100 rectangles. Comment on the performance of the mean of the estimates of total area from SRS from the table above and compare it to the mean of the estimates from the judgement samples. Which is a better estimate of the true population total area and explain why one estimate should be better than another?

**Solution:**

Based on the above two graphs, it appears that for SRS, the range of the estimated mean is closer to the population mean as compared to the estimated mean from judgement sample, thus proving to be better. This is not necessary always the case. Since SRS chooses the samples randomly, the range of the mean obtained from SRS would lie anywhere between the minimum and maximum values. SRS gives equal weightage to all the possible samples from the population. There would be very less scenarios that would lead to the estimation to be closer to the extremes and large number of possible samples that would lead to estimation closer to the mean. The notable observation here is the bell-shaped curve formed in case of SRS with the center of the bell very close if not equal to the population mean, which can be said as the mean of the sampling distribution from SRS is equal to the population mean given enough random samples are picked (Central Limit Theorem).

In case of single sample selection, judgement sample may sometimes perform better depending on how good the judgement was. But mostly the SRS will perform better since its randomly picking from the available sample. If the population is large, then judgement sample has very less chances of performing better as it would become hard to analyze the population and pick up the right sample. Also, if there is any change in available population data (for example, during exploratory data analysis (EDA), it was found that some outlier entries should not be used as they were result of an error), judgmental sampling may require to re-analyze the data before sample selection making it more time consuming and an expensive process in term of effort required. Whereas in case of SRS, random samples can be pulled directly without any re-analysis or any similar change.

So, SRS should be given preference over judgement sampling.

**Exercise 6**: What is the minimum possible value of estimated total area that could be obtained from a SRS of size 10 from the rectangular population? What is the maximum possible value of estimated total area that could be obtained from a SRS of size 10?

**Solution**

estimated total area = (Total Sample Area / Sample Size) \* 100

To get in minimum possible area, the rectangles selected should be 10 rectangles without replacement having minimum area. On analyzing the rectangles, we found that there are 16 rectangles with area 1, so to obtain the minimum area, the sample should have all the 10 rectangles from this set of 16 rectangles (ID as 1, 2, 3, 4, 5, 8, 9, 10, 11, 16, 19, 20, 21, 44, 50 and 64) which has a unit area. So the minimum possible value of estimated total area = 10 / 10 \* 100 = 100.

Similarly, to get in minimum possible area, the rectangles selected should be 10 rectangles without replacement having maximum area. On analyzing the rectangles, we found that maximum rectangle area is 18 and there are 5 rectangles having area as 18. The second largest value of area is 16 and there are 10 such rectangles. So to have sample with maximum area, all the 5 rectangles (ID as 25, 57, 70, 76, 96) having area as 18 should be in the sample with the remaining 5 from the 10 rectangles (ID as 13, 31, 41, 47, 51, 55, 60, 73, 90, 93) having the area of 16. So the maximum possible value of estimated total area = (18 \* 5 + 16 \* 5) / 10 \* 100 = 1700

Minimum possible area from sample of size 10: 100

Maximum possible area from sample of size 10: 1700

**Exercise 7**: Is it possible that the estimated total area from an SRS could be worse than your judgement sample estimated total area? Explain.

**Solution:**

Yes, it is possible since the random selection of samples in SRS makes possible to pick any possible sample which includes samples resulting in minimum and maximum areas as well. In judgement samples, since sample is picked selectively based on the understanding and knowledge of the data, its not very likely that the judgmental sample would result in values close to maximum or minimum values.

**Exercise 8**: Is it likely that the estimated total area from an SRS could be worse than your judgement sample estimated total area? Explain.

**Solution:**

It depends. The closer the judgement sample estimate is to the actual mean of the population, the greater is the likelihood of the estimated total from the SRS to be worse than the judgement sample and vice versa. In order to get the most accurate estimation of the total area, the best way is to take multiple random samples from the population and then compute the sample mean. This mean value should result in the most accurate estimation of the total area. In case of a single random sample, since it can pick any sample from the worst to the best, it may perform worse then the judgement sample and that likelihood increases with the accuracy of judgement sample