**Lab 3. Sampling distributions and stratified sampling**

**MSDS 6370**

## Hari Narayan Sanadhya

**Objective:**

* For the student to learn about sampling distributions for simple random samples without replacement (SRSWOR).
* For the student to gain insight into stratified samples.

**Introduction**

The topic of the reading material for Asynchronous Lecture 3 was a discussion of sampling distributions. In this lab, we continue to study sampling distributions. In addition we consider a particular type of stratified sampling design.

**Estimating population mean and variance from a sample**

Two characteristics of a population *yi* of size *N* are its mean and its variance. The mean is the average of the population values and the variance indicates the spread of those values. Often we see the square root of the variance , called the standard deviation. The formulas we use today to calculate these are

and .

If we take a sample of size *n* from our population, we use the sample mean to estimate the population mean and also use the sample variance to estimate the population variance as follows:

and .

**Population data**

In this assignment, you will use a small population of 8 taxpayers. Data is in Table 1 and gives the actual income of the taxpayers. An auditor does not have the actual data for the 8 taxpayers and plans to take a sample without replacement of size 2 to estimate the actual income of these taxpayers. He does know the reported income of all 8 taxpayers as shown in Exercise 3, Table 3. The objective of the audit is to estimate the mean of the difference between the actual and reported incomes of the 8 taxpayers.

**Table 1:** Actual income for 8 taxpayers

|  |  |
| --- | --- |
| **Taxpayer number** | **Actual Income (thousands of dollars)** |
| 1 | 60 |
| 2 | 72 |
| 3 | 68 |
| 4 | 94 |
| 5 | 90 |
| 6 | 102 |
| 7 | 116 |
| 8 | 130 |

**Exercise 1**

1. Systematically list all 28 possible samples of size n = 2 from the taxpayer population of size 8 in Table 1. For each sample, calculate the mean and the standard deviation and enter these in Results Table 1 in the Results Tables Section at the end of this document.

2. Calculate the population mean and standard deviation of actual income of all 8 taxpayers in Table 1 and compare them to the sampling mean and the standard deviation of the sampling distribution of the mean.

3. Show that the mean of the standard deviation s of samples in Results Table 1 is not equal to the standard deviation of the sampling distribution of the mean and explain why there is a difference.

4. Construct a histogram of the 28 sample estimates of the mean in Results Table 1. This is the *sampling distribution of the mean*. How does the shape of this histogram compare to the shape of a normal distribution?

5. Construct a histogram of the 28 sample estimates of the standard deviation. How does the shape of this histogram compare to the shape of a normal distribution.

**Exercise 2**

The auditors later discover a 9th missing taxpayer with the data below:

**Table 2: Taxpayer 9 Data**

|  |  |
| --- | --- |
| **Taxpayer number** | **Actual Income (thousands of dollars)** |
| 9 | 200 |

They decide they must include Taxpayer 9 in their sample because of her large income. In this case, we say that Taxpayer 9 is included with certainty or with a probability of selection equal to 1. If we include Taxpayer 9 in our analysis and the data from Exercise 1, we essentially have a stratified sample of size 3 where 2 units are selected from the original 8 and the third is selected with certainty. Exercise 2 examines the effect on the estimator of accounting for the stratification in estimating the population mean.

By adding Taxpayer 9 the population size is N = 9. Since the auditors want Taxpayer 9 to be in the sample with certainty, we essentially have 2 strata, Stratum 1 of size N1 =8 with the original 8 Taxpayers and Stratum 2 of size N2 =1 with a single taxpayer, Taxpayer 9. Essentially, we have a stratified sample of size 3 where 2 are selected from Stratum 1 with the original 8 Taxpayers and 1 from Stratum 2 with only Taxpayer 9 whose actual income is 200. The mean of a sample of size 2 from Stratum 1 is and the mean of a sample of size 1 from Stratum 2 is , which equals 200 since Stratum 2 has only 1 Taxpayer. Therefore, the estimate of the population mean incorporating the stratification is

.

1. Use the estimates of the mean actual income for Taxpayer 9 and all possible samples of size 2 from the original 8 taxpayers that you entered into Results Table 1 and calculate all possible estimates of . Enter these in Results Table 2 shown in the Results Tables Section at the end of this document.

2. Calculate the sampling mean and standard deviation in Results Table 2 and compare them to the population mean income and standard deviation of the 9 taxpayers.

3. Construct a histogram of the 28 sample estimates of the mean in Results Table 2. How does the shape of this histogram compare to the shape of a normal distribution?

**Exercise 3**

You are given the data below on reported income for 8 taxpayers.

**Ex3**. If you hypothesize that the difference in mean actual and reported income is > 10, what percent of the time would you confirm your hypothesis if you used a large number of samples of size 2 of actual and reported income to test your hypothesis? Explain you answer.

**Table 3: Reported Income 8 Taxpayers**

|  |  |
| --- | --- |
| **Taxpayer number** | **Reported Income (thousands of dollars)** |
| 1 | 50 |
| 2 | 56 |
| 3 | 66 |
| 4 | 76 |
| 5 | 90 |
| 6 | 100 |
| 7 | 112 |
| 8 | 110 |

## **Solutions:**

## **Exercise 1**

1. Refer to Results Table 1 below

**Results Table 1: 8 Taxpayers**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample number** | **2 of 8 taxpayers**  **in sample** | **mean of actual income of 2 taxpayers** | **Estimate of std. dev.** |
| 1 | 1, 2 | 66 | 8.485281 |
| 2 | 1, 3 | 64 | 5.656854 |
| 3 | 1, 4 | 77 | 24.04163 |
| 4 | 1, 5 | 75 | 21.2132 |
| 5 | 1, 6 | 81 | 29.69848 |
| 6 | 1, 7 | 88 | 39.59798 |
| 7 | 1, 8 | 95 | 49.49747 |
| 8 | 2, 3 | 70 | 2.828427 |
| 9 | 2, 4 | 83 | 15.55635 |
| 10 | 2, 5 | 81 | 12.72792 |
| 11 | 2, 6 | 87 | 21.2132 |
| 12 | 2, 7 | 94 | 31.1127 |
| 13 | 2, 8 | 101 | 41.01219 |
| 14 | 3, 4 | 81 | 18.38478 |
| 15 | 3, 5 | 79 | 15.55635 |
| 16 | 3, 6 | 85 | 24.04163 |
| 17 | 3, 7 | 92 | 33.94113 |
| 18 | 3, 8 | 99 | 43.84062 |
| 19 | 4, 5 | 92 | 2.828427 |
| 20 | 4, 6 | 98 | 5.656854 |
| 21 | 4, 7 | 105 | 15.55635 |
| 22 | 4, 8 | 112 | 25.45584 |
| 23 | 5, 6 | 96 | 8.485281 |
| 24 | 5, 7 | 103 | 18.38478 |
| 25 | 5, 8 | 110 | 28.28427 |
| 26 | 6, 7 | 109 | 9.899495 |
| 27 | 6, 8 | 116 | 19.79899 |
| 28 | 7, 8 | 123 | 9.899495 |

1. Population mean and standard deviation is 91.5 and 22.7101299 respectively.   
   Using the formula of standard error for sampling distribution i.e.

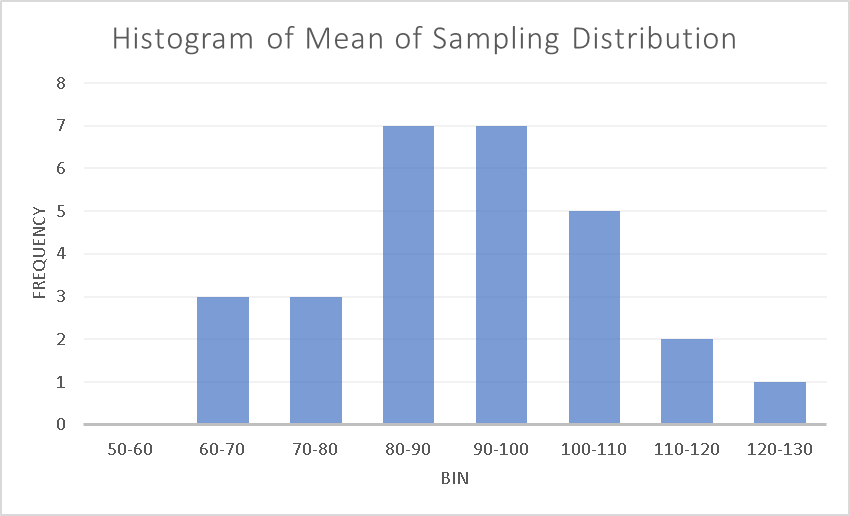


the result of the standard error that we obtain matches with standard deviation obtained when the samples are drawn with replacement and the order of the samples matter.

|  |  |
| --- | --- |
| Mean | Std Dev |
| Population | |
| 91.5 | 22.7101299 |
| Sampling Dist Without Replacement and order doesn’t matter (28 Samples) | |
| 91.5 | 14.8672699 |
| Sampling Dist With Replacement and order doesn’t matter (36 Samples) | |
| 91.5 | 16.92713141 |
| Sampling Dist With Replacement and order does matter (64 Samples) | |
| 91.5 | 16.05848685 |
| Sampling Dist using the formula for mean and std deviation | |
| 91.5 | 16.05848685 |

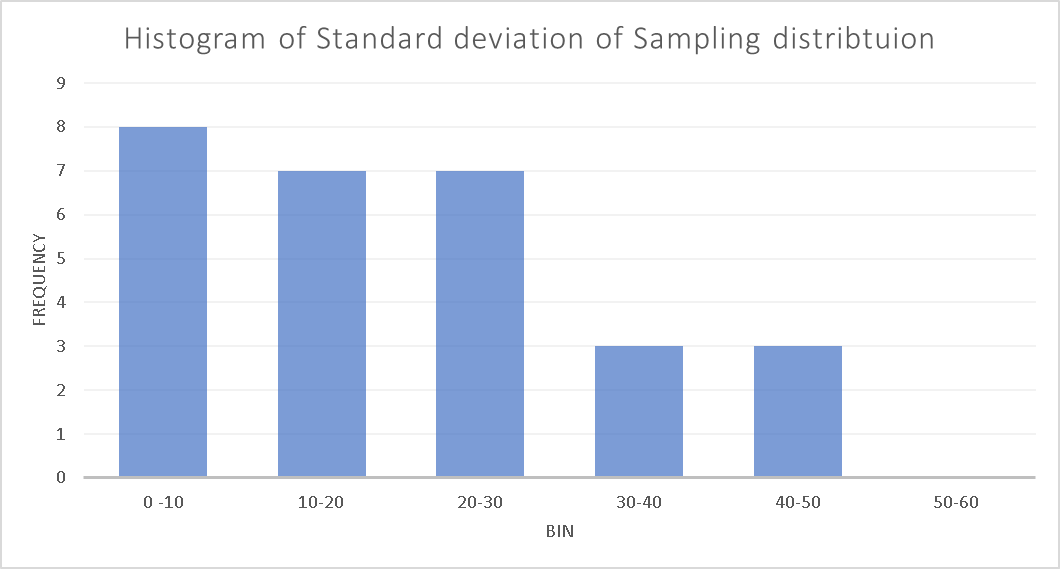
For the sample distribution obtained from the 28 samples, the mean and the standard deviation obtained is 91.5 and 14.87. The mean of the sampling distribution matches with the mean of population but the standard deviation is much smaller in case of sampling distribution.

1. The mean of standard deviations of all the 28 samples in the sampling distribution comes out to be 20.80914242 which is much greater than the standard deviation of the sampling distribution (14.87). This is because standard deviation of the sampling distribution measure of square root of variance of the means of sampling distribution. Its not the mean of the standard deviations of all the samples but is instead the variability in the mean of the samples.
2. The histogram created using Excel is as below:



The histogram resembles with the shape of the normal distribution (bell shape curve with the center of the bell at the mean which here is 91.5). So, the sampling distribution is normally distributed.

1. The histogram of the standard deviation of the 28 samples is not normal distributed but is rather right skewed with most of the observation in the initial range (0-30) maximum in 0-10 range. This makes sense because it means that most of the observations are close to one another and as we see from the histogram of means of sampling distribution, the means of samples are close to the center (population mean) hence forming the bell curve.



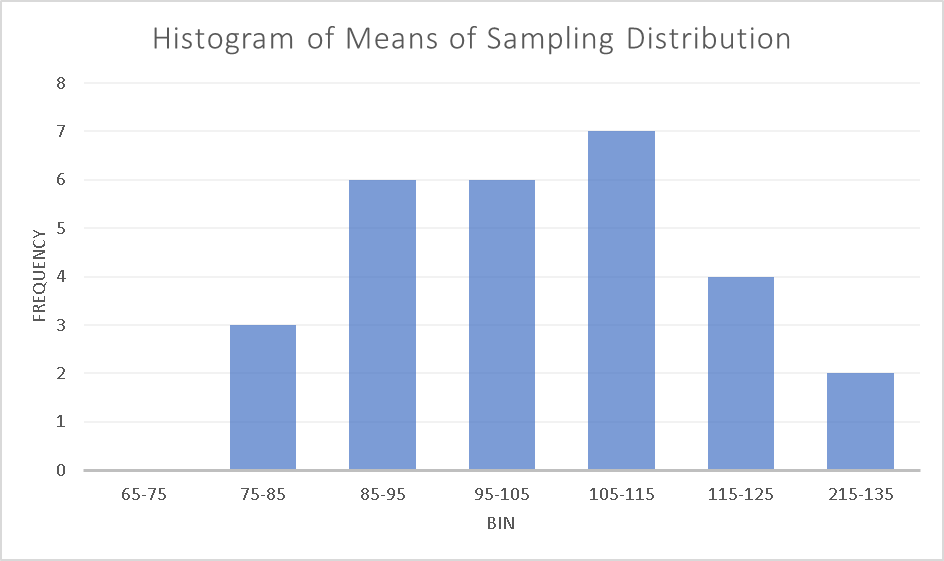
## **Exercise 2**

1. Refer to Results table 2 below

**Results Table 2: 9 Taxpayers**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample number** | **2 of 8 taxpayers**  **in Stratum 1** | **mean of actual income of 2 taxpayers from Stratum 1,** | **Estimate of population mean of actual income for 9 taxpayers** |
| 1 | 1, 2 | 66 | 80.88889 |
| 2 | 1, 3 | 64 | 79.11111 |
| 3 | 1, 4 | 77 | 90.66667 |
| 4 | 1, 5 | 75 | 88.88889 |
| 5 | 1, 6 | 81 | 94.22222 |
| 6 | 1, 7 | 88 | 100.4444 |
| 7 | 1, 8 | 95 | 106.6667 |
| 8 | 2, 3 | 70 | 84.44444 |
| 9 | 2, 4 | 83 | 96 |
| 10 | 2, 5 | 81 | 94.22222 |
| 11 | 2, 6 | 87 | 99.55556 |
| 12 | 2, 7 | 94 | 105.7778 |
| 13 | 2, 8 | 101 | 112 |
| 14 | 3, 4 | 81 | 94.22222 |
| 15 | 3, 5 | 79 | 92.44444 |
| 16 | 3, 6 | 85 | 97.77778 |
| 17 | 3, 7 | 92 | 104 |
| 18 | 3, 8 | 99 | 110.2222 |
| 19 | 4, 5 | 92 | 104 |
| 20 | 4, 6 | 98 | 109.3333 |
| 21 | 4, 7 | 105 | 115.5556 |
| 22 | 4, 8 | 112 | 121.7778 |
| 23 | 5, 6 | 96 | 107.5556 |
| 24 | 5, 7 | 103 | 113.7778 |
| 25 | 5, 8 | 110 | 120 |
| 26 | 6, 7 | 109 | 119.1111 |
| 27 | 6, 8 | 116 | 125.3333 |
| 28 | 7, 8 | 123 | 131.5556 |

1. The population mean income and standard deviation of the 9 taxpayers comes out to be 103.5555556 and 40.26333074 respectively. The mean and standard deviation obtained from the sampling distribution is 103.5555556 and 13.21535102 respectively. The mean of the sampling distribution is same as the population mean but the standard deviation obtained from the sapling distribution is very small as compared to population standard deviation (approximately 33.33 % of the population standard deviation).
2. The Histogram of sampling distribution is close to normal distribution. It shows a little left skewness and this is because of the outlier which is the 9th sample and small sample size (3) and always selecting sample 9. If the samples were drawn using SRS instead of stratified, the skewness may not have been visible and the samples would have been almost perfectly normal distributed if not otherwise.



## **Exercise 3**

To find the percentage of times when the difference in mean actual and reported income is > 10, we can use the power of central limit theorem. To confirm the hypothesis, since we can use large number of samples of size 2 of actual and reported income and compute their difference in means. But here the two populations are not independent so we cannot draw independent samples from the two populations. In fact they have data corresponding to the same 8 taxpayers. So instead we will combine them and then check if the difference between the two incomes is greater than 10 or not. So, our population would be: -

|  |  |  |  |
| --- | --- | --- | --- |
| **Taxpayer number** | **Actual Income (thousands of dollars)** | **Reported Income (thousands of dollars)** | **Difference between actual and reported income** |
| 1 | 60 | 50 | 10 |
| 2 | 72 | 56 | 16 |
| 3 | 68 | 66 | 2 |
| 4 | 94 | 76 | 18 |
| 5 | 90 | 90 | 0 |
| 6 | 102 | 100 | 2 |
| 7 | 116 | 112 | 4 |
| 8 | 130 | 110 | 20 |

Now if we generate the 28 possible samples from this,

|  |  |  |
| --- | --- | --- |
| **Sample Number** | **2 of 8 taxpayers  in sample** | **Mean of Difference between actual and reported income of taxpayers in sample** |
| 1 | 1, 2 | 13 |
| 2 | 1, 3 | 6 |
| 3 | 1, 4 | 14 |
| 4 | 1, 5 | 5 |
| 5 | 1, 6 | 6 |
| 6 | 1, 7 | 7 |
| 7 | 1, 8 | 15 |
| 8 | 2, 3 | 9 |
| 9 | 2, 4 | 17 |
| 10 | 2, 5 | 8 |
| 11 | 2, 6 | 9 |
| 12 | 2, 7 | 10 |
| 13 | 2, 8 | 18 |
| 14 | 3, 4 | 10 |
| 15 | 3, 5 | 1 |
| 16 | 3, 6 | 2 |
| 17 | 3, 7 | 3 |
| 18 | 3, 8 | 11 |
| 19 | 4, 5 | 9 |
| 20 | 4, 6 | 10 |
| 21 | 4, 7 | 11 |
| 22 | 4, 8 | 19 |
| 23 | 5, 6 | 1 |
| 24 | 5, 7 | 2 |
| 25 | 5, 8 | 10 |
| 26 | 6, 7 | 3 |
| 27 | 6, 8 | 11 |
| 28 | 7, 8 | 12 |

Out of the 28 cases, we have 10 samples for which the mean of difference between actual and reported income of taxpayers in sample is greater than 10. So the percent of the time the hypothesis is true is 10/28 = 35.7143%