**MSDS 6379**

**Lab 2: Weighted Sample Designs**

**Submitted by Hari Narayan Sanadhya**

**Objectives:**

* Introduce the student to weighted sample designs.
* Allow the student to understand the advantages of weighted vs. no non-weighted sample designs

You are a consultant and you are asked to analyze residents' responses to the question: "How much should your town spend on education per pupil." You are asked to do this for two neighboring towns. You decide to take samples of residents using email. The local internet provider has given you data about internet usage per month by person in the towns. You expect your response rate will depend on how often a person uses the internet. You collect data for 25 residents in each town using email based on a random sample across all residents (see samplesData sheet in lab2DatForStudents.xlsx file). Complete the Exercises below. Please label your answers clearly and give concise and thorough explanations where appropriate.

**Exercise1**: Calculate the mean and standard deviation of the email sample responses for each town based on simple random sampling.

**Exercise2**: Comment on whether these are good estimates of the means for the towns.

**Exercise3**: Use another way to calculate an estimate of the population means that corrects for selection bias due to internet usage and explain why you chose this method.

**Exercise4**: Calculate the means for data from an in-person survey done last year (see faceSurveyData sheet in lab2DatForStudents.xlsm file). Explain why you do not have to calculate a weighted sample.

**Exercise5**: Compare the means you calculated in the email samples to the data from the in-person survey done last year.

**Exercise6**: Create histograms using your email samples (Hint: You will have to define histogram ranges). Compare the histograms for the two towns.

**Exercise7**: Create histograms using the in-person data from last year.

**Exercise8**: Compare the email sample and the in-person survey results and histograms for the two towns. Is your email sample or the in-person survey a better estimate of the population means? Explain.

**Submit Answers Below**

**Exercise1**: - Measurement of the sample statistic values

|  |  |  |
| --- | --- | --- |
|  | Town1 | Town2 |
| Mean | 50.56 | 37.44 |
| Standard Deviation | 11.33240192 | 6.916646586 |

**Exercise2**: -No, these statistics are not the good estimates of the population parameters. Since the above statistic is from SRS, the sample obtained could be any of the possible samples. Using just the single sample instead of sampling distribution to get estimates the parameters is very likely to give result far from the actual. Also since the sampling does not consider the internet usage, the result is biased thus not a good estimate. Persons with high internet usage are more likely to respond as they will likely be checking their emails more frequently. Since the survey method used is via email, non-representativeness would always be there as only email/internet users will be able to respond and that too, all the email receivers would be responding. So, this unrepresentativeness and non-response from the receivers needs to be accommodated as well.

**Exercise3**: -

The method chosen to correct the selection bias due to internet usage is by use of Weights that will down weight persons with high internet usage and up weight persons with low internet usage. As per the problem description, the response rate on the email depends on how often a person uses the internet. This means that the probability of selection is not equal and cannot be determined given the data. This is a case of non-probability sample. We have been provided with the per month internet usage of each user. Since our sample has information regarding the internet usage but the probability of selection of users based on the internet usage cannot be determined based on the given information, we cannot calculate the weights using the probability formula (wi = 1/πi). But since we know that the probability of selection is directly proportional to the internet usage, we can compute the weights by using this relative probability.

**For Town 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Town 1** | | | | |
| **Person ID** | **Response ($k)** | **Internet Usage** | **Weight** | **Response \* Weight** |
| 3147 | 39 | 22 | 0.045454545 | 1.772727273 |
| 442 | 46 | 22 | 0.045454545 | 2.090909091 |
| 1284 | 52 | 19 | 0.052631579 | 2.736842105 |
| 2043 | 65 | 19 | 0.052631579 | 3.421052632 |
| 2569 | 59 | 20 | 0.05 | 2.95 |
| 179 | 32 | 22 | 0.045454545 | 1.454545455 |
| 1672 | 49 | 20 | 0.05 | 2.45 |
| 1741 | 58 | 19 | 0.052631579 | 3.052631579 |
| 2306 | 62 | 20 | 0.05 | 3.1 |
| 3009 | 33 | 18 | 0.055555556 | 1.833333333 |
| 1810 | 57 | 22 | 0.045454545 | 2.590909091 |
| 511 | 46 | 18 | 0.055555556 | 2.555555556 |
| 1905 | 56 | 21 | 0.047619048 | 2.666666667 |
| 1409 | 59 | 18 | 0.055555556 | 3.277777778 |
| 1146 | 56 | 21 | 0.047619048 | 2.666666667 |
| 3078 | 28 | 22 | 0.045454545 | 1.272727273 |
| 2638 | 61 | 21 | 0.047619048 | 2.904761905 |
| 1478 | 58 | 16 | 0.0625 | 3.625 |
| 41 | 40 | 23 | 0.043478261 | 1.739130435 |
| 1215 | 55 | 19 | 0.052631579 | 2.894736842 |
| 1974 | 57 | 20 | 0.05 | 2.85 |
| 373 | 34 | 22 | 0.045454545 | 1.545454545 |
| 2375 | 68 | 20 | 0.05 | 3.4 |
| 110 | 38 | 18 | 0.055555556 | 2.111111111 |
| 1879 | 56 | 20 | 0.05 | 2.8 |
|  |  | **Total** | **1.254311214** | **63.76253934** |

Estimated Mean = =

**For Town 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Town 2** | | | | |
| **Person ID** | **Response ($k)** | **Internet Usage** | **Weight** | **Response \* Weight** |
| 1774 | 38 | 14 | 0.071428571 | 2.714285714 |
| 958 | 26 | 12 | 0.083333333 | 2.166666667 |
| 290 | 48 | 12 | 0.083333333 | 4 |
| 1900 | 33 | 15 | 0.066666667 | 2.2 |
| 769 | 39 | 15 | 0.066666667 | 2.6 |
| 1535 | 34 | 14 | 0.071428571 | 2.428571429 |
| 2893 | 38 | 11 | 0.090909091 | 3.454545455 |
| 2226 | 46 | 16 | 0.0625 | 2.875 |
| 895 | 33 | 17 | 0.058823529 | 1.941176471 |
| 530 | 33 | 12 | 0.083333333 | 2.75 |
| 2202 | 43 | 17 | 0.058823529 | 2.529411765 |
| 2831 | 48 | 15 | 0.066666667 | 3.2 |
| 1661 | 32 | 18 | 0.055555556 | 1.777777778 |
| 416 | 40 | 16 | 0.0625 | 2.5 |
| 1837 | 29 | 13 | 0.076923077 | 2.230769231 |
| 655 | 36 | 15 | 0.066666667 | 2.4 |
| 2654 | 45 | 15 | 0.066666667 | 3 |
| 2139 | 28 | 15 | 0.066666667 | 1.866666667 |
| 2014 | 37 | 19 | 0.052631579 | 1.947368421 |
| 2352 | 49 | 10 | 0.1 | 4.9 |
| 832 | 36 | 13 | 0.076923077 | 2.769230769 |
| 2591 | 40 | 16 | 0.0625 | 2.5 |
| 1598 | 25 | 13 | 0.076923077 | 1.923076923 |
| 353 | 45 | 15 | 0.066666667 | 3 |
| 2768 | 35 | 14 | 0.071428571 | 2.5 |
|  |  | **Total** | **1.765964896** | **66.17454729** |

Estimated Mean = =

**Exercise4**: - Measurement of the sample statistic values

|  |  |  |
| --- | --- | --- |
|  | Town1 | Town2 |
| Mean | 51.8 | 39.84 |
| Standard Deviation | 9.844626284 | 9.254188 |

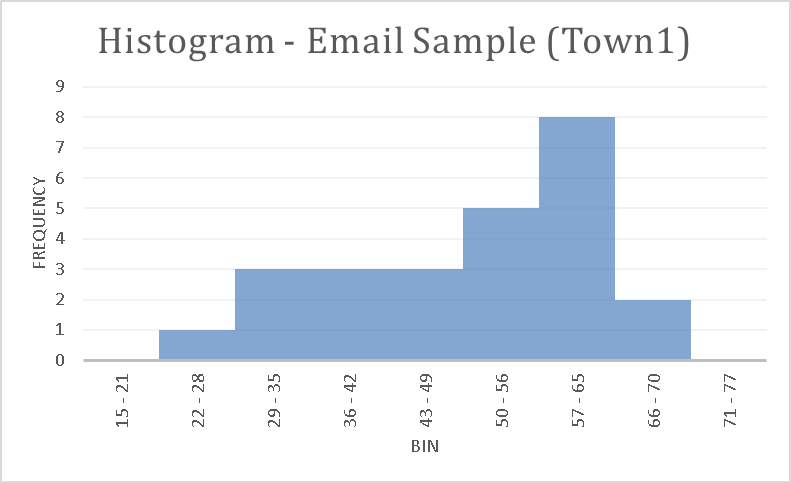
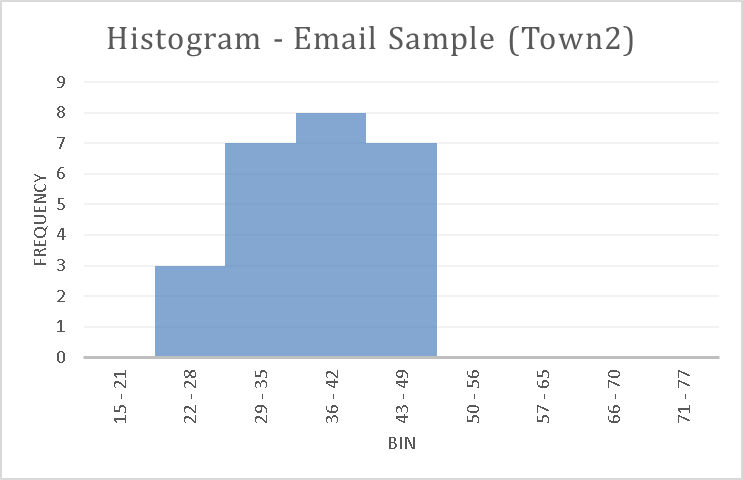
Since the survey method is a face to face survey, there is no need to weight this by the internet usage. If the sample selected is random and no bias is introduced while collecting survey data (like selective picking of the person to reach for in person survey), there is no need to apply any weights in this case as all the person chosen will have equal probability of selection,

**Exercise5**: -

For town 1, the means obtained from SRS, weight SRS and in-person survey data collected last year are 50.56, 50,535 and 51.8 respectively. The means from SRS and weighted SRS are almost similar but differs a little from the mean calculated from the last years in person survey data though the difference is not statistically significant (p-value of 0.6587 for 95% confidence).

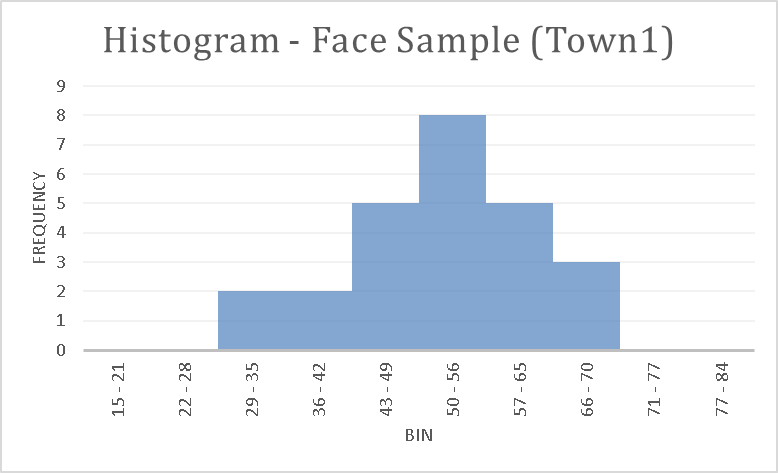
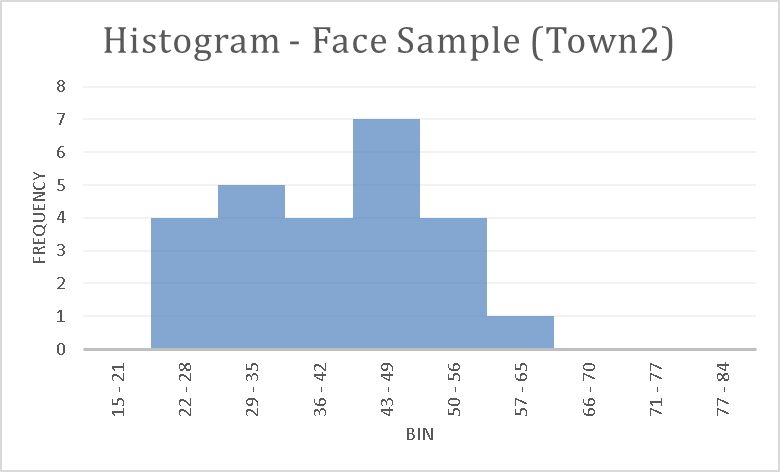
For town 2, the means obtained from SRS, weight SRS and in-person survey data collected last year are 37.44, 39.472 and 39.84 respectively. The means from last years in person survey data and weighted SRS are almost similar but differs a little from the mean calculated from SRS though the difference is not statistically significant (p-value of 0.8463 for difference of means for 2-sided t-test).

**Exercise6**: -Created using Excel

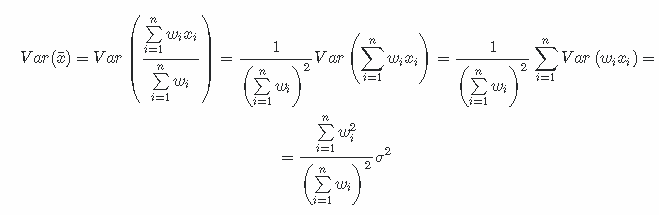
The Histograms shows that for town 1, spread of the samples collected is wider then for town 2. Town 1 sample is left skewed and has large variance as compared to the sample from town 2 which is a little right skewed and has small variance. Since this is a random sample, this will not always happen.

**Exercise7**: - Created Using Excel

**Exercise8**: -

For weighted means, using the formula



Source - <https://math.stackexchange.com/questions/320441/standard-deviation-of-the-weighted-mean>

We computed the standard deviation of samples for town 1 and town 2 obtained from the email survey data. We found that the standard deviation of the weighted means for town 1 and town 2 is 2.275 and 1.4 respectively.

From the town 1 histograms above, the email sample looks left skewed whereas, the in-person survey sample looks not skewed. When looked at the means obtained from both of them, they are almost the same but have different variance. The standard deviation from SRS, weighted SRS and in person survey sample are 11.33, 2.275 and 9.844. Since the means are not statistically different, the comparison was done using the standard deviation. The weighted means have the least value and hence is the best ones of the three.

From the town 2 histograms above, both the email sample and the in-person survey sample appear to have a little right skewness. When looked at the means obtained from both of them, they are almost the same but have different variance. The standard deviation from SRS, weighted SRS and in person survey sample are 6.91, 1.4 and 9.254. Since the means are not statistically different, the comparison was done using the standard deviation. The weighted means have the least value and hence is the best ones of the three.

Though we have mentioned that results from the weighted means looks the best, but because of the issue of non-representativeness in the population (access to internet/emails is must for email survey), the data from in person survey may be closer to the actual population then email survey. But for in person survey too, there may be bias present due to selective picking of the persons to reach for survey by the surveyor or refusal by the individual to participate in the survey, the information about which is missing. So, looks like there is some sort of bias in both the cases so would vote for the one which is giving the better result here – email survey sample by weighted means.