ST 531 Project

On Ping Lai

December 08, 2022

1 Introduction

A popular topic while discussing women's right is low income females. Ideally, if men and women have equal right, then the proportion of low income in terms of sex should be the same (i.e. equal amount of male and female in low income). However, for a long time, there were always significant more low income female then male. In 21st century, does women right improved? Is there still more low income female than male in the United States?

2 Methodology

To answer the question, I found this data set about the recipients of Supplemental Security Income Program (SSI) in 2010. SSI is a nationwide federal program that provides monthly payment to low income and disability people. It is a census that each data point is distinguished by the STATE, SEX, and AGE. We maybe able to use the proportion of female receiving SSI payment to study if there is more low income female in the US. The data set consists 50 states, District of Columbia, and Northern Mariana Islands(Guam); 13 age groups " ¡05 ", " 05-12", " 13-17", " 18-21", " 22-25" ," 26-29", " 30-39", " 40-49", " 50-59", " 60-64", " 65-69", " 70-74", " 75+ "; 2 sexes "M" and "F" ; and the number of recipients that have certain characteristics. For example, we can read the first row of the data point as: " there are 84 female recipients from Alaska that is age below 5". Thus, there are $52 * 2 * 13 = 1352$ data point in total.

However, most of the time we are unable to obtain population data. In this paper, I would like to pretend that I do not know any population information and estimate the population proportion of female receiving SSI payments using randomized sampling. Under this setting, the target population is people living in the US; the sampled population is all the SSI recipients in 2010; the Sampling Unit and Observation Unit are SSI recipients; the response of interest is the gender of recipients.

A two-stage-cluster sampling will be sufficient since a group that identified by ages and states usually has a large within-groups variance. In practice, we usually like to begin with selecting the states then sample the age group of the people in that selected state.

In Stage 1, I randomly choose 5 states as our two stages-cluster from these 52 districts using R code sample(state, 5). Here I defined "state" as a vector consists the name of 52 districts in the US. This guarantees each state is chosen in equal probability. The output returns : (Idaho, Alaska, Louisiana, Tennessee, West Virginia). Thus, we have $N = 52$ and $n = 5$. In Stage 2 , using similar method, I randomly choose 4 of the age groups from 13 age groups. The output is (" 18-21", " 70-74", " <05 ", " 50-59"). Thus, we have $Mi = 13$ and $mi = 4$ for each $i = 1, ..., 5$. Then, I obtained the following tables:

3 Results

The first table is the number of female recipients in a certain State and age group.

	ID <int>	AK <int>	LA <int>	TN <int>	WV <int>
18-21	760	242	3490	2993	1130
70-74	640	428	5443	5239	1953
<05	326	84	2387	1962	492
50-59	2933	1425	19531	21317	10594
Total	4659	2179	30851	31511	14169

The second table is the number of total recipients in a certain State and age group.

	ID <int>	AK <int>	LA <int>	TN <int>	WV <int>
18-21	1785	630	8662	7538	2944
70-74	933	721	7591	7684	3119
<05	812	212	5723	4514	1255
50-59	4706	2534	32865	36184	18224
Total	8236	4097	54841	55920	25542

Here we are estimating the population proportion by the Ratio Estimator. It is because we are estimating the population proportion, which is the mean of a binomial setting (female = 1 , male = 0). Besides, we also assumed we do not know any information about the population and we can only rely on the data that we randomly draw in the previous section.

3.1 Point Estimate

The Point Estimate of a two-stage- clustering is:

$$\hat{p} = \frac{\sum_{i=1}^{n} M_i * \hat{p_i}}{\sum_{i=1}^{n} M_i} = 0.5556656$$

which each $p_i = \{0.5656872, 0.5318526, 0.5625536, 0.5635014, 0.5547334\}$ is calculated by the sum all female counts from the sampled age groups in the same state divided by the total counts from the sampled age groups in the same state.

3.2 Variance and Standard Error

The variance of the estimate is:

$$\hat{\mathrm{V}} = \frac{1}{\bar{M}^2}(1 - \frac{n}{N})\frac{s_r^2}{n} + \frac{1}{nN\bar{M}} \sum M_i^2 (1 - \frac{m_i}{M_i})\frac{s_i^2}{m_i} = 0.01427046$$

where $s_r^2 = \dfrac{\sum(M_i\hat{p_i} - M_i\hat{p})^2}{n-1} = 0.03283697$

and $s_i^2 = \dfrac{m_i\hat{p_i}(1-\hat{p_i})}{m_i - 1} = \{0.3275803, 0.3319806, 0.3281161, 0.3279568, 0.3293390\}$

Therefore, the standard error is 0.119459.

3.3 Confidence Interval

The 95% confidence interval is given by:

$$\hat{p} \pm z^* \sqrt{\hat{\mathrm{V}}} = (0.3215259, 0.7898054)$$

We are 95% confident that between 0.3215259 and 0.7898054 of the lower income people in US is female.

4 Discussion

Since the data set is a census, we can find that the true proportion of female recipient is 0.547, which is within the confidence interval. Indeed, our estimate 0.556 is very close to the true proportion. Nevertheless, although age (e.g. kids cannot go to work or older people more experienced and get higher pay) and state(e.g. technology industries in CA and WA, Wall Street in NY) are both important factors of income level, there are still many other confounding factors. A stratified sampling based on years of education maybe another method to give a good estimation to answer the research question.

Regardless if the estimation is close to the true value, there are non-sampling errors while using the proportion of female SSI recipients maybe both overcoverage and undercoverage in some level. First, SSI also release payment to disability people. Although it is arguably that low-income and disability has a high correlation, it may still include a group of people that is not within our target population. Second, there maybe recipients who were not low-income nor disabled people receiving the payments since recipients lied or the system wrongly approved the payment. In addition, not all low-income people noticed the existence of SSI program and thus did not apply for the payment. Therefore, they were not in the data set, which undercoverage the target population. Besides, the SSI also encounters self-selection biased, since the program requires application. Due to the societal pressure, it might be more difficult for male to admit they are in low income.

On the other hand,

5 Conclusion

This paper performed a two-stage-cluster sampling based on a census data. The estimated proportion 0.556 is very similar to the true proportion 0.547. This indicates that the methodology of two-stage-clustering is not the major issue. What is more concerning is the level of representative. Even the SSI data set itself is a census, it is still a subset of low-income population. The confounding variables and non-sampling errors would affect the proportion of low income female.

6 Reference

Data set:

https://catalog.data.gov/dataset/supplemental-security-income-ssi-recipients-in-each-state-by-sex-and-age-december-2010

Social Security Administration Office of Research, Evaluation, and Statistics (2010). Supplemental Security Income (SSI) Recipients by Geographic Area, Sex, Age, Eligibility, and Diagnostic Group, 2010 Data. Data.gov. https://www.ssa.gov/policy/docs/data/ssi-2010/SSI-2010-User-Manual.pdf