Investigation on income inequality of the Islands
STAT3003 Midterm B - Survey Report

Zheng Weijia (William, 1155124322)
Department of Statistics, The Chinese University of Hong Kong

May 13, 2021

Abstract

Following the survey plan completed one month ago in April, while adopting its Mathematical model and overall structure of the designing of sampling method, we improved some technical points in that plan to measure the income inequality of the Islands towns better. In this report, we will show the specific mothod we chose, the survey results and their interpretations as well.

1 Introduction

Income inequality problem is drawing worldwide attention, it directly affects the life quality of many of us, especially the poor, which is in line with our intuition and daily experience in the avenues and streets of our city of Hong Kong, which is a region famous for its acutely huge gap between rich and poor. On the other hand, as stated by the Washington Post, income inequality also hurts economic growth, especially high inequality in rich nations. Awaring of the importance of studying in this topic, we decide to measure the level of income inequality for each town of the Islands as an economy of a certain scale.

Based on the survey plan (i.e., the Midterm A part), we improved the original entropy measurement. We do not ues the Gini coefficient because Gini coefficient does not have additive property.

2 Mathematical Model to Measure Income Inequality

Suppose a town has a population of n. And the total wealth of this town is W. Denote the average income as $\bar{w} = \frac{W}{n}$. For every person i (i ranges from 1 to n) in this town, denote the wealth he possesses is w_i, define

$$y_i := \frac{w_i}{\bar{w}} \ln \frac{w_i}{\bar{w}}, \ \forall i.$$

We adpot the Theil index Mathematical model, (applied in OECD, the Organisation for Economic Co-operation and Development) which is defined as

$$T := \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{\bar{w}} \ln \frac{w_i}{\bar{w}}.$$

This measurement is improved from the simple entropy measurement I proposed in part A, which treats income inequality as the uncertainty of whose pocket, among those all possible n

people, will a dollar go into, while treating everyone's possibility of earning it as proportional to his current wealth. In addition, this T index has an extra property that it valuates regions irrespective of its extent, which means the problem that our original model encourages more population is overcomed. From T index's definition, it is a μ-typed metric, with special case $T = 0$ stands for all money are evenly distributed and any other (higher) value represents a higher level of disproportion.

3 Core Survey Method

We adopted an two-stage cluster sampling. Imagine a town with T houses. (1) To form an large cluster, choose an integer N and all house whose house numbers are congruent to modulo N are grouped into a same cluster. Therefore N also stands for the number of clusters in the population. (2) To do the first-stage sampling, we SRS $n = 4$ from the N clusters. For each selected cluster, we denote the number of elements (number of valid people) inside as M_i. (3) To do the second-stage sampling, we SRS m_i from the M_i elements.

The unbiased point μ estimator formulae for two-stage cluster sampling is

$$\hat{\mu} = \hat{T} = \frac{N}{Mn}\sum_{i=1}^{n}\hat{y}_i,\ \hat{y}_i = M_i\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}.$$

And the formulae of estimated variance is

$$\widehat{Var(\hat{\mu})} = \frac{1}{M^2}N(N-n)\frac{1}{n}{\hat{\sigma_c}}^2 + \frac{1}{M^2}\frac{N}{n}\sum_{i=1}^{n}M_i(M_i-m_i)\frac{1}{m_i}{\hat{\sigma_i}}^2.$$

Where the ${\hat{\sigma_c}}^2$ is the sample variance of the estimated cluster totals and ${\hat{\sigma_i}}^2$ is the sample variance inside cluster i.

Hence an appropriate $100(1-\alpha)\%$ C.I. can be given by

$$(\hat{\mu} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\mu})}}\ ).$$

Recall that we issued in the A part that when using ordinary cluster sampling, the clusters are tend to be too large. Not only to be more efficient, but also base on the fact that the population is large and supposely many elements inside a cluster are similar, we chose to go with two-stage cluster sampling method.

4 Case Demonstration

4.1 Always Subtract the Number of Kids

Another problem issued by A part is about the kids. Kids' having no money should not be counted for income inequality. In our formulae, M and M_i are related with number of kids. The latter can be easily handled because we observe every element of cluser i. We can get a very accurate estimated M, the number of all non-kid (older than 12 years old), by subtracting (1) the number of all 0 to 5 years old babies, which can be obtained by inspecting the born record in Town Hall, and (2) the number of school students, whose ages are 5 to 12 strictly.

4.2 Case Example: Hofn

Hofn is a northern town in the northernmost island of the three. It has a total population of 2143, with total number of houses being 1055.

After eliminating the number of preschooler (96) and school students (246), we have the number of valid people being

$$M = 2143 - 96 - 246 = 1801.$$

Take $N = 150$, we SRS $n = 4$ numbers between 1 and 150 inclusively, result to be 7, 20, 82, 101. Let $m_1 = m_2 = m_3 = m_4 = 4$, conducting the sampling, we have

$$M_1 = 11, M_2 = 13, M_3 = 12, M_4 = 14.$$

With

$$\bar{Y_1} = -0.1926, \bar{Y_2} = 0.6950, \bar{Y_3} = 0.8072, \bar{Y_4} = 0.3352.$$

Using the formulae

$$\hat{Y_i} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} Y_{ij} = M_i \bar{Y_i},$$

we can have

$$\hat{Y_1} = -2.1186, \hat{Y_2} = 9.0355, \hat{Y_3} = 9.6859, \hat{Y_4} = 4.6925.$$

From $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y_i}$, we can have $\bar{Y} = 5.3238$. And then the point estimate of τ is

$$\hat{\tau} = N \frac{1}{n} \sum_{i=1}^{n} \hat{Y_i} = \frac{150}{4} \cdot 5.3238 = 798.5736.$$

Hence

$$\hat{\mu} = \frac{1}{M} \hat{\tau} = \frac{1}{1801} \cdot 798.5736 = 0.4434.$$

Note that

$$\hat{\sigma_1}^2 = 0.0103, \hat{\sigma_2}^2 = 0.4692, \hat{\sigma_3}^2 = 1.1882, \hat{\sigma_4}^2 = 0.5203.$$

Using the formulae

$$\hat{\sigma_c}^2 = \frac{\sum_{i=1}^{n} (\hat{Y_i} - \bar{Y})^2}{n - 1},$$

we have $\hat{\sigma_c}^2 = 29.5308$. Then

$$\widehat{Var(\hat{\mu})} = \frac{1}{M^2} N(N-n) \frac{1}{n} \hat{\sigma_c}^2 + \frac{1}{M^2} \frac{N}{n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{1}{m_i} \hat{\sigma_i}^2 = 0.050547.$$

Therefore a 90% C.I. for the T index of Hofn can be given by:

$$(\hat{\mu} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\mu})}}\,) = (0.4434 \pm 1.64 \cdot \sqrt{0.050547}\,) = (0.4434 \pm 0.3687).$$

5 Survey Results

5.1 Results for Other Towns

We selected some other towns and applied the above survey method, the selected towns are: Talu, Nelson, Takazaki and Valais.

Talu:
$\hat{\mu} = 0.2965$, and $\widehat{Var(\hat{\mu})} = 0.0127$.
A 90% C.I. can be $(0.2965 \pm 1.64 \cdot \sqrt{0.0127}) = (0.2965 \pm 0.1848)$.

Nelson:
$\hat{\mu} = 0.4128$, and $\widehat{Var(\hat{\mu})} = 0.0515$.
A 90% C.I. can be $(0.4128 \pm 1.64 \cdot \sqrt{0.0515}) = (0.4128 \pm 0.3722)$.

Takazaki:
$\hat{\mu} = 0.1969$, and $\widehat{Var(\hat{\mu})} = 0.04239$.
A 90% C.I. can be $(0.1969 \pm 1.64 \cdot \sqrt{0.04239}) = (0, 0.5346)$.

Valais:
$\hat{\mu} = 0.60204$, and $\widehat{Var(\hat{\mu})} = 0.03545$.
A 90% C.I. can be $(0.60204 \pm 1.64 \cdot \sqrt{0.03545}) = (0.60204 \pm 0.3088)$.

5.2 Interpretation of Survey Results

	T index point estimate	Total Population
Hofn	0.4434	2143
Talu	0.2965	1077
Nelson	0.4128	551
Valais	0.6020	530
Takazaki	0.1969	357

Above is the table of our survey results, comparing with the population of the towns. We can see that T index is functioning well, not affected by population. From this table, together with the previous results, we are confident to draw a conclusion that Valais is very likely to have serious wealth inequality problem, while Takazaki and Talu are likely to be good in this regard.

5.3 Difficulties and Problems