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.

As suggested by the survey plan (i.e., the Midterm A part), we adopted the entropy index so called "H index" in the plan. We do not ues the Gini coefficient because this Mathematical model has much better additive property hence it is possible for us to "estimate" it as a τ-typed metric from a fraction of the population.

2 Mathematical Model to Measure Income Inequality

Suppose a town has a population of n. And the total wealth of this town is W. For every person i (i ranges from 1 to n) in this town, denote the wealth he possesses is w_i, define

$$y_i := -\frac{w_i}{W} \log_n \frac{w_i}{W}, \ \forall i.$$

Then the H index of this town is defined as

$$H := \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} -\frac{w_i}{W} \log_n \frac{w_i}{W}.$$

This measurement treats income inequality as the uncertainty (more academically, entropy) 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.

From H index's definition, it is a τ-typed metric, with value ranges from 0 to 1. 1 stands for the case that the income is evenly distributed to everyone, while 0 stands for the case that someone evil and greedy is holding all the money.

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 τ estimator formulae for two-stage cluster sampling is

$$\hat{\tau} = \hat{H} = N\frac{1}{n}\sum_{i=1}^{n}\hat{y_i},\ \hat{y_i} = M_i\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}.$$

And the formulae of estimated variance is

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

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

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

Recall that in the A part we issued two problems, one of them is 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

$$\bar{Y_1} = 1.42792 \cdot 10^{-4}, \bar{Y_2} = 5.39019 \cdot 10^{-4}, \bar{Y_3} = 5.35803 \cdot 10^{-4}, \bar{Y_4} = 3.85439 \cdot 10^{-4}.$$

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

$$\hat{Y_1} = 5.71169 \cdot 10^{-4}, \hat{Y_2} = 2.156076 \cdot 10^{-3}, \hat{Y_3} = 2.14321 \cdot 10^{-3}, \hat{Y_4} = 1.541755 \cdot 10^{-3}.$$

From $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y_i}$, we can have $\bar{Y} = 1.60305 \cdot 10^{-3}$. And then the point estimate is

$$\hat{\tau} = \hat{H} = N \frac{1}{n} \sum_{i=1}^{n} \hat{Y_i} = \frac{150}{4} \cdot 6.41221 \cdot 10^{-3} = 0.24046.$$

Note that

$$\hat{\sigma_1}^2 = 2.8518 \cdot 10^{-8}, \hat{\sigma_2}^2 = 9.85337 \cdot 10^{-8}, \hat{\sigma_3}^2 = 1.40172 \cdot 10^{-7}, \hat{\sigma_4}^2 = 1.1658 \cdot 10^{-7}.$$

Using the formulae

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

we have $\hat{\sigma_c}^2 = 5.55382 \cdot 10^{-7}$. Then

$$\widehat{Var(\hat{\tau})} = N(N-n)\frac{1}{n}\hat{\sigma_c}^2 + \frac{N}{n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{1}{m_i} \hat{\sigma_i}^2 = 3.44855 \cdot 10^{-3}.$$

Therefore a 95% C.I. for the H index of Hofn can be given by:

$$(\hat{\tau} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\tau})}}\ ) = (0.24046 \pm 1.96 \cdot \sqrt{3.44855 \cdot 10^{-3}}\ ) = (0.24046 \pm 0.1151\ ).$$

5 Survey Results

5.1 Results for All Towns

5.2 Interpretation of Survey Results