Investigation on income inequality of the Islands
STAT3003 Midterm A - Survey Plan

ZHENG Weijia (William, 1155124322)

April 6, 2021

Abstract

In this report paper, I suggested to measure the income inequality situation of towns in Islands as a τ-type metric using the way one measuring the entropy of a random variable, instead of using the complex Gini coefficient. By denying the feasibilities of other sampling methods, I suggested to use a cluster sampling with a systematic sampling way to form clusters. Though this approach and the sampling methods remain many flaws, I did a small scale trial sampling and gave the result generated.

1 Introduction and Notations

Income inequality continues to receive world-wide attention. The most famous method to measure it to use the Gini Index proposed by Gini. Gini index gives a number inside the interval $[0, 1]$, the more the index close to 1, the worse the inequality problem is.

Assume we will investigate the situation of income inequality for the scale of a town (e.g., Hofn, Vardo and Takazaki are all towns).

Then denote the population of the town as n, note that this is given.

Denote the total amount of wealth of the town as M, which is also given by consulting the Hall of the town.

Denote the average amount of wealth as $\mu = \frac{M}{n}$.

For the i-th individual (i ranges from 1 to n), denote the his individual amount of wealth as m_i, note that $\sum_{i=1}^{n} m_i = M$.

Then if we follow the definition of Gini index, we need to calculate

$$G = \frac{1}{2n^2\mu} \sum_{j=1}^{n} \sum_{i=1}^{n} |m_i - m_j|.$$

But this is troublesome because it is hard to estimate the G of a town properly from a sample of smaller size. Hence we switch to another method of measuring the inequality.

2 Mathematical Model

2.1 Derived from the Concept of Entropy

As said above, hindered by the complex way to calculate the Gini index out and noted that the Islands provide the information of the total amount of a town and every individual's. We can then

understand the concept of "income euqality" from the perspective of the difference between "every person's ability (as a possibility) to gain wealth".

For the i-th person, we regard the ratio between his own wealth and the total wealth, i.e., $p_i = \frac{m_i}{M}$ of the town as his "possibility of gaining a unit of wealth".

Note that $\sum_{i=1}^{n} p_i = 1$, hence the p_i forms a proper probability distribution. Note that the entropy

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_n p_i$$

actually measures the inequality of the town economy. Hence define it as the "H index" of the economy.

When the town economy is the most fair, i.e., everyone has the same probability of earning money: $p_i = \frac{1}{n}, \forall i = 1, 2, \ldots n$, then

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_n p_i = -\sum_{i=1}^{n} \frac{1}{n} \log_n \frac{1}{n} = 1.$$

When the economy is extremly inequal, i.e., all the money go to one person, then the entropy will be 0.

The following is a simple example of the application of this Mathematical model. Consider a economy with only 3 people, each income is 10, 20, 15. And another economy with 3 people with each income 3, 7, 40. The first economy has $H_1 = 0.9656$ and the second economy with $H_2 = 0.5666$, which conforms our intuition.

Note that the H index is a τ-type metric, so all the knowledge for estimating τ can be applied properly.

3 Sampling Method

3.1 Not Using Auxiliary Data

Recall that we need to estimate a τ-type metric, and the metric t_i which we need to get from each individual (denote $t_i = -p_i \log_n p_i$) is a deterministic function of m_i because $p_i = \frac{m_i}{M}$. Since m_i is extremly easy to obtain, we will not apply the "using auxiliary data" trick.

3.2 Not Using SRS

If we apply SRS, then we need a whole list which contains every one inside the town and randomly choose some from it as our sample.

Although we know the population number, but all people are grouped in houses, and we cannot know how many people are inside each house, since we cannot have such a whole list to randomly choosing elements from. Therefore we do not choose SRS.

3.3 Not Forming Clusters Geographically

This is an empirical decision from my investigating on some towns. I observed that the neighbourhoods of a person is likely to have similar economic status with him. In other words, people sharing similar economic status are likely to gather together geographically as a cluster. Hence I will not use Cluster Sampling under such way of forming clusters.

3.4 Our Sampling Method

We will adopt Cluster Sampling by forming clusters in a special way, more explicitly, in a systematic sampling manner.

First, we select a proper number s as the stepsize, and r.

We will classify all the people whose house number modulus s equals r.

Note that there are $N = \frac{T}{s}$ clusters in totals, where T is the number of NON-EMPTY houses in the town.

We choose N_c clusters from the N clusters by SRS.

A good point I want to emphasis here is that, each cluster we construct is geographically scattered all over the town, hence they are likely to have small difference between clusters and larger difference inside a cluster, which is good.

Then, for every chosen cluster $i, i = 1, 2, \ldots N_c$ we go through all the houses and the people inside those houses and collect data from every element(person).

Using formulae

$$\hat{\tau} = \hat{H} = N\frac{1}{N_c}\sum_{i=1}^{N_c} Y_i,$$

with $Y_i = \sum_{j=1}^{M_i} t_{ij}$, where $M_i, i = 1, 2, \ldots N_c$ is the number of elements (people) inside the i-th chosen cluster. We have the point estimate.

Recall that $\hat{\tau}$ is unbiased.

Using the formulae

$$\hat{Var}(\ \hat{\tau}\ ) = N^2(1 - \frac{N_c}{N})\frac{\hat{\sigma_c^2}}{N_c},$$

we can have an estimated variance, where the $\hat{\sigma_c^2}$ is the sample variance of the cluster totals, i.e., the variance of those $Y_i's$, with $i = 1, 2, \ldots N_c$.

Note that we need to let $N_c > 1$ to make sure the $\hat{\sigma_c^2}$ is making proper sense.

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

$$\hat{\tau} \pm t_{N_c-1,1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\ \hat{\tau}\ )}.$$

4 Preliminary Survey Result

We select the town Valais to be under investigation. Valais contains 255 houses.

And we let the stepsize being 17, and choose $N_c = 4$ (this is to let the $t_{N_c-1,0.975}$ be at a low level) number from the 1 to 17 to get started. Using a calculator's random number function, I obtained: 2, 6, 14 and 17.

Hence I need to investigate about $4 \times 15 = 60$ houses.

To sum up, $N_c = 4, N = 17$.

And from the investigating, $Y_1 = 0.020314809, Y_2 = 0.028761294, Y_3 = 0.029845762, Y_4 = 0.02049111$.

The point estimate is

$$\hat{\tau} = \frac{17}{4}(Y_1 + Y_2 + Y_3 + Y_4) = 0.44129.$$

And the estimated variance should be

$$\hat{Var}(\ \hat{\tau}\ ) = 1.47 \times 10^{-3}.$$

With $t_{3,0.975} = 3.182446$, therefore a 95% C.I. for the H index of Valais is

$$0.4413 \pm 0.1220.$$

.

This seems to be varying a lot, but I think as it is a metric about the income inequality of a whole town, it is understandable and I believe there are ways to eliminate the variance width.

5 Where to Imporve

5.1 Problems of Mathematical Model

A easy point to be noted is that this entropy model encourages an economy to scaling up with its wealth distribution structure unchanged. While the Gini coefficient regard the above two things' income inequality level as equal. Therefore the more population a town have, it is very likely that the less inequality it will be under my model.

5.2 Problems of Survey Method

To be honest, there are a few problems in my designing of my survey method.

5.2.1 Kids Won't Have Money

Empirically speaking, kids, especially those under 15 years old, will not have money, while those between 15 and 19 are likely to have very few money. This is caused because they are not able to earn money yet, which is no business with income inequality, but in my survey I did not split kids out and count they having no money into the situation of income inequality.

5.2.2 Too Few and Too Large Clusters

In the preliminary survey, I only divide the whole town into 15 clusters. Comparing with the examples in the lecture notes, my designing is too few. The clusters should be larger and this is also good for eliminating the t value.