STAT410 Project

Sampling Designs of Estimating the Average Price of Hostels in Tokyo

Jiajun Zhang: 301327368

LingXiang Zou: 301289420

Tan Tian: 301293751

April 3rd, 2019

Introduction

Maximizing the use of the travel budget is often the most important thing that travellers consider. Nowadays, instead of staying in a hotel, there is also a large number of travellers tend to choose hostels as their accommodation because of the low cost and decent quality. Hostels are no longer just for lone backpackers and group of students, but it is also becoming popular among older adults, families and even business traveller who only have a tight budget.

As all of our group members have never been to Japan before, we are planning to visit Tokyo, the capital city of Japan. For accommodations, we consider price as our priority as we do not have plenty of budgets and we are interested in how the hostels help us saving money than other lodgings. In this project, we use different sampling design strategies to estimate average pricing of hostels in Tokyo, then compare it to the average pricing of hotels and Airbnb to see if it is more affordable for us. The data set we use contained a total of 127 different hostels in Tokyo, and corresponding information for each hostel. After variables selection, we choose price per night as our variable of interest and distance from the city center as our auxiliary variable.

Methods

1.Simple Random Sampling

Simple random sampling (without replacement) is the basic sampling design that selects units from the target population, and each unit is equally likely to be selected. In our case, the maximum difference between our estimate and the true price we can accept is 250 yen. By formulas, the desired sample size is 43, but to achieve an accurate estimate, we use our sample size n=50 in this design and other designs as well in order to maintain consistency of this research. We randomly generated 50 samples from the population to calculate the sample mean. Then, reiterate the process 10000 times to get the average value of the 10000 results which is sample mean.

Regression Estimation

In reality, the further the distance from the city center the lower the price should be. Since our auxiliary variable is the distance from the Tokyo center, there is potentially a strong linear relationship between the variables. Even though, our scatter plot has the trend of going down, but not a strong linear relationship. To obtain the linear regression estimator of our

model, we use SRS to get 50 samples and calculate the slope and y-intercept of the least squares regression line that fits the sample data. Once we have the estimator, $\widehat{\mu L} = a + b\,\mu x$, we can substitute the population mean of our auxiliary variable μx to get our sample mean. Again, reiterate the process 10000 times, and take the average value of those sample means.

2.Stratified Random Sampling

Stratified sampling involves the use of "stratum" which are the subsets of the target population. In our case, the population is partitioned into four strata based on the distance in kilometres (km) from Tokyo center. Each stratum has equal size interval of 5 in between 0 to 20 kms. By briefly looking at the points of price and distance in Figure 1, we can tell there are some variabilities of hostels pricing within 10 kms from downtown Tokyo, especially in the range of 5km. It makes sense in real-life situations where the housing prices in the downtown city tend to be more expensive, considered a hotel price in downtown Vancouver. However, stratified sampling handles such case well because although the units within one stratum have a large variance, determining a desired number of stratified sample size for each stratum can be representative of the population. This also explains the case where we want to use optimum allocation to determine sample sizes within each stratum.

Nonetheless, the units within the same stratum are homogeneous, meaning that the prices of hostels within each distance interval are close. Because of this, we expect that stratified random sampling would have a more accurate estimation than SRS. Since our strata are different in size, we use both proportional and optimum allocations to obtain the sizes we want to sample within each stratum. After 10000 iterations of calculating the stratified sample mean by using SRS with both allocations, the MSE of optimum allocation performs better than proportional allocation. This is what we expected because the optimum allocation allocates with the lowest variability within each stratum which minimizes the variance and MSE.

3.Stratified with Unequal Probability Sampling

The unequal probability sampling method is used in our survey design while we like to observe more hostels with shorter distances to Tokyo center. This makes sense sometimes in a place where urbanization really achieved. A similar example of such city is Beijing where it has multiple so-called "rings roads" surround its city center, as the map shown in Figure 2. In

this case, we can think of at the edge of each ring road, it is more developed than other places within that ring road since it is closer to the center. Then, using our auxiliary variable distance, we first assign the probability to each of them in a sense that a shorter distance gets a higher probability to be selected. Simply, we can name such design as simple random sampling with replacement with selection probabilities proportional to distance. However, this method does not work quite well for us while the distribution of hostels is not as what we expected. Also, the plot suggests that there is a high density of hostels in the middle range of distances. Similarly, we use this method within each the stratum we just created, assuming that the auxiliary variable distance tends to get more important for hostels with shorter distances in strata. However, as shown in Figure 3, getting a poor result of MSE using HH, HT, and GUPE estimators after simulations lead us to believe that the assumption does not hold well.

Results

	Mean	MSE	Approximate Bias	Acutal CI coverage
Simple Random Sampling (SRS)	2769.091	13059.210	-0.750	0.937
Regression Estimation	2762.894	12164.453	-6.948	0.929
Stratified SRS, Proportional Allocation	2769.042	10755.629	-0.799	0.939
Stratifed SRS, Optimum Allocation	2769.828	9446.465	-0.013	0.941
Population	2769.841			

From the mean comparison, all these designs return good estimates, so we cannot conclude which design performs the best. However, based on other simulation results, the stratified random sampling with optimum allocation returns us a smallest mean square error, relatively low bias, and highest probability confidence interval coverage. Since we used a nominal 95% confidence interval on each estimator with different strategies, all of the true coverage probabilities did not meet 95%. The reason they are different is that the sampling distribution is asymmetric which is not normally distributed. In our actual CI coverage comparison, stratified random sampling with optimum allocation has the closest probability coverage to our nominal confidence interval. As well as MSE and bias comparisons, this design seems to have the least errors with high accuracy. This is because the units within each stratum are very similar, and tend to have a better estimation. Whereas, SRS has a more stable

estimation if the data is more spread and regression estimation has more accurate result if the variables are highly correlated. Lastly, the stratified random sampling with proportional allocation and regression estimation tend to do a better job than SRS, when considered their MSE's. As a result, stratified random sampling with optimum allocation is the most reliable effective design to estimate the population mean in our topic.

Discussion

The result shows that our best estimation with the lowest mean square error is stratified random sampling with optimum allocation. Assuming that in reality, all our group members are the data collectors, and we are going to Tokyo to collect the information of each hostel and calculate the average price per night. We might want to use stratified sampling because every group member can be responsible for one stratum, and then we can randomly select a small sample size of hostels in each stratum. A large portion of travel time between strata can be saved by doing this. Relatively speaking, stratified sampling is not necessarily more expensive than simple random sampling which means that it will not cost the collectors more time or money, but can lead to a better result. In fact, our dataset is not informative enough, except for the auxiliary variable, distance to the city center, in our designs. There are many other potential auxiliary variables associated with the price such as the size of the room, the age of the hostels, and etc. However, we do not have those data in our dataset; if those variables are included in the regression model, it might effectively improve the results of multiple regression estimator. That way, the regression strategy might be more desirable than stratified with optimum allocation.

Next year, in 2020, the thirty-second Olympics game is going to take place in Tokyo. The amount of expected foreign visitors will increase significantly. According to the result above, the dataset we found online which is about hotel price in Tokyo and the Airbnb website, we find that if tourists consider hostels as their choice of accommodation, it will save them approximately 30 CAD per night compared to hotels. Comparing to Airbnb, it will save about 100 CAD per night. Thus, we suggest that hostels can significantly benefit people who are planning to travel to Tokyo on a budget or looking to meet new people with different nationalities and backgrounds. However, for those who want to have more privacy or have extra money to spend, hostels might not be a good choice.

Reference

Mohn, T. (2013). Tight Travel Budget? Try a Hostel. New York Times (1923-Current File), p. B8.

Japan Hotels Datasets - TokyoHotels - ISL Resource Catalogue. (n.d.). Retrieved from http://islcatalog.ics.forth.gr/de/dataset/japan-hotels-datasets/resource/56c89557-8494-4686-866a-4555fdc6308d

Vacation Rentals, Homes, Experiences & Places. (n.d.). Retrieved from https://www.airbnb.com/

Thompson, S. K. (2012). *Sampling*. Hoboken, NJ: Wiley.

Ring roads of Beijing. (2019, March 28). Retrieved from https://en.wikipedia.org/wiki/Ring_roads_of_Beijing

Appendix of Figures

Figure 1

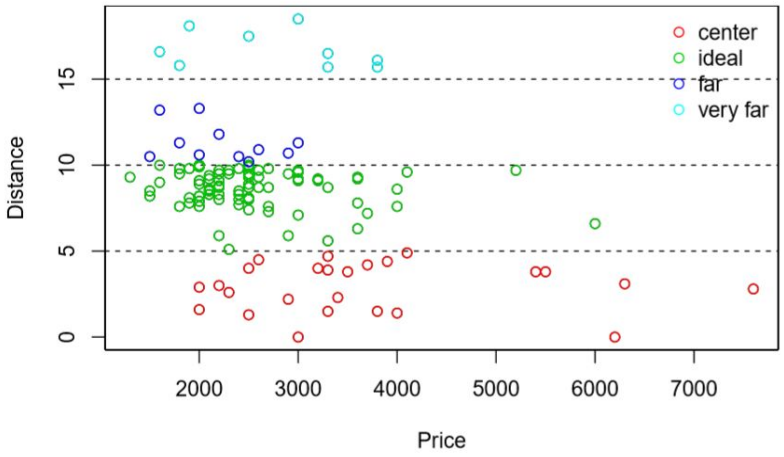

Figure 2

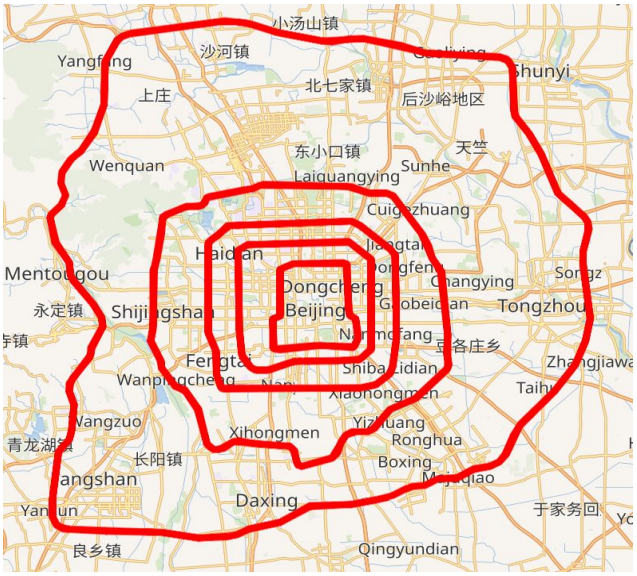

Figure 3

MSE of Unequal Probability Sampling with Replacement

```
##       SRS        HH        HT      GUPE
## 13059.21 60680.79 83226.80 27444.25
```

MSE of Stratified Unequal Probability Sampling with Replacement

```
##       SRS        HH        HT      GUPE
## 13059.21 24500.02 42766.13 15606.86
```