STAT344 Group Project

Professor: Lang Wu Date:

Nov 13, 2023

Group leader:
Wenhui Bao, 59773879
Songru(Suri) Chang, 45792728

Group members:
Zhihao Wu, 56834187; Yuge Xing, 44264299,
Songru Chang, 45792728; Weiqi Cai, 14104210

Roles:
R code: Wenhui Bao, Songru Chang
Analysis: Wenhui Bao, Weiqi Cai, Songru Chang, Zhihao Wu, Yuge Xing
Writing: Songru Chang, Zhihao Wu, Yuge Xing

PART I

Airbnb, the pioneer of the new "sharing economy" and the originator of a business concept that has become a global phenomenon, is an innovative American company that allows people to rent out their vacant rooms or homes to travellers. Airbnb has grown rapidly since its inception, expanding the possibilities for guests and hosts to travel, presenting a more unique and personal way of experiencing the world. The United States has the most significant amount of international tourism revenue globally, and accommodation revenue accounts for the most significant part of tourism revenue. So, as travel enthusiasts, we have the motivation to research the prices of Airbnb in New York, USA. This investigation selects relevant data from a website named Kaggle and uses SRS (simple random sampling) and STR (stratified sampling) to estimate the average price of Airbnb in America through the price of Airbnb in New York.

The study takes the dataset from the Kaggle website that contains around 49,000 New York Airbnb and their prices, and we treat them as the population. We use two sampling methods: simple random sampling (SRS) and stratified sampling (STR, proportional allocation with respect to neighbourhood groups). There are two parameters in the research. The first variable is continuous, which illustrates the mean price of all Airbnb in New York. The other variable is binary, which takes “1”s if the price is higher than or equal to the mean price of the stratified sample and “0” s if the price is smaller than the stratified sample.

After comparing the two sampling methods, we found that simple random sampling has many advantages. It is easy for researchers to carry out the sample, being able to select each Airbnb into the sample with equal chance. However, it may not involve extreme cases with smaller counts, and the sample may not be spread out enough to epitomize the whole population. Also, more significant errors may be obtained from the simple random sample with the same sample size compared to other sampling methods. For stratified sampling, it provides a preciser mean estimator when subpopulations are heterogeneous and very different from each other. Also, we can reduce survey costs by simplifying data collection. Moreover, the stratified sample can ensure that the sample represents all groups, even when some groups have relatively smaller sample sizes. However, there are also some cons to stratified sampling. For instance, we require a standard scheme for every stratum so that every member of the population fits into their strata. In other words, sufficient information is needed for assigning the members into strata, and some individuals may be hard to be classified.

In order to compare sampling methods, we choose two samples with the same size via two different methods. For simple random sampling, we randomly sample 1200 Airbnbs from the total population with size N = 48,884. Due to the fact that one of the neighbourhood regions named ‘Staten Island’ only contains 373 units, which is a relatively small count of Airbnb’s prices in the total population, the sample size has to reach 1200 such that individuals in all levels of neighbourhood groups are chosen into the sample, and we want to find the mean price of Airbnb that can represent the whole population. For stratified sampling, we divide the data into five stratum with respect to 5 different neighbourhood groups, including Bronx, Brooklyn, Manhattan, Queens and Staten Island. We choose the sample stratum sizes using proportional allocation, based on population stratum sizes and the formula: $\frac{n_h}{N_H} = \frac{n}{N}$.

.

Under simple random sampling, we analysis the data based on the equation: $\bar{y}_s = \frac{\sum_{i\in S} y_i}{n}$, SE($\bar{y}_s$) =$\frac{s}{\sqrt{n}}$, $95\%\ CI = \bar{y}_s \pm 1.96 * \text{SE}(\bar{y}_s)$, where n is the sample size 1200, y_i's are the Airbnb prices, $\bar{y}_s$ is the sample mean of the prices, s is the sample standard deviation. It is obvious that the average price of New York Airbnb is $155.58, and the standard error is $9.17. The 95% confidence interval is between $137.61 to $173.55, which implies that when conducting the study 100 times, there are approximately 95 times that the actual population average price of New York Airbnb stays between $137.61 and $173.55. Also, under stratified sampling, we analysis the data based on the following formulas: $\bar{y}_{str} = \sum_{h=1}^{5} \frac{N_H}{N} \overline{y_{sh}}$, SE($\bar{y}_{str}$)=$\sqrt{\sum_{h=1}^{5} \frac{N_h^2}{N^2}(1 - \frac{n_h}{N_h}) \frac{S_{sh}^2}{n_h}}$, $95\%\ CI = \bar{y}_{str} \pm 1.96 * \text{SE}(\bar{y}_{str})$, where $\bar{y}_{str}$ is the mean price of the stratified sample, $\overline{y_{sh}}$ is the sample mean price of each strata, N_h is the population strata size of each neighborhood group, n_h is the sample strata size of each neighborhood group, S_{sh}^2 is the sample variance of each strata. The average price of New York Airbnb is $151.26, and the standard error is $5.01. The 95% confidence interval is between $141.44 to $161.08, which indicates that by conducting the study 100 times, there are approximately 95 times that the average population price of New York Airbnb stays between $141.44 and $161.08. Therefore, stratified sampling generates a narrower sample average price range of the true population average price with a smaller standard error compared to the simple random sampling.

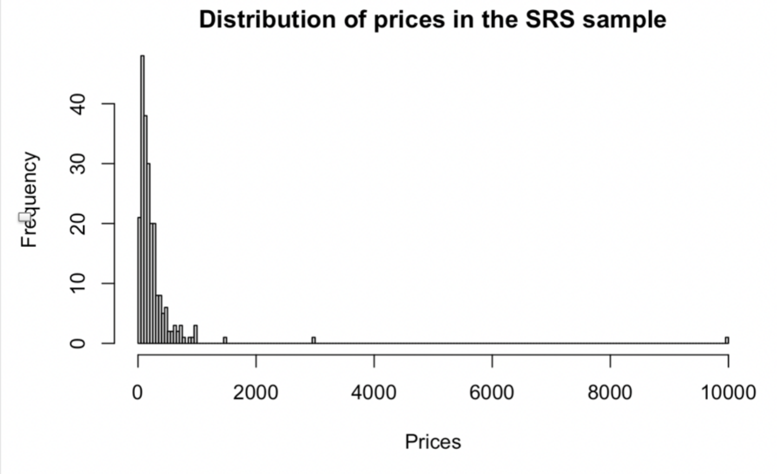


Figure1. Distribution of Airbnb's prices in SRS sampling

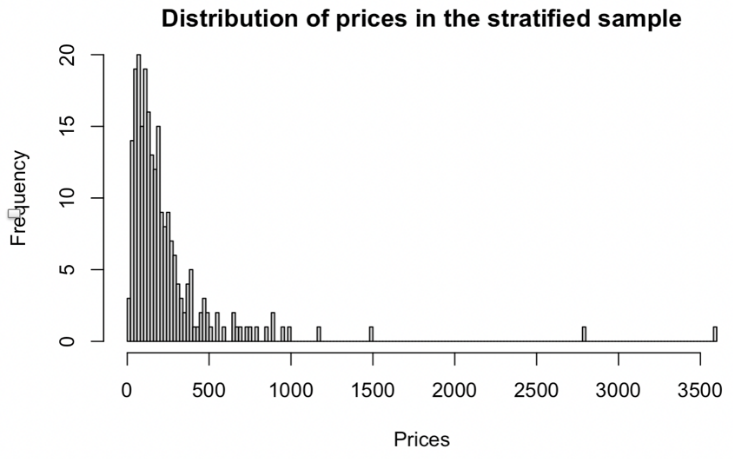


Figure2. Distribution of Airbnb's prices in STR sampling

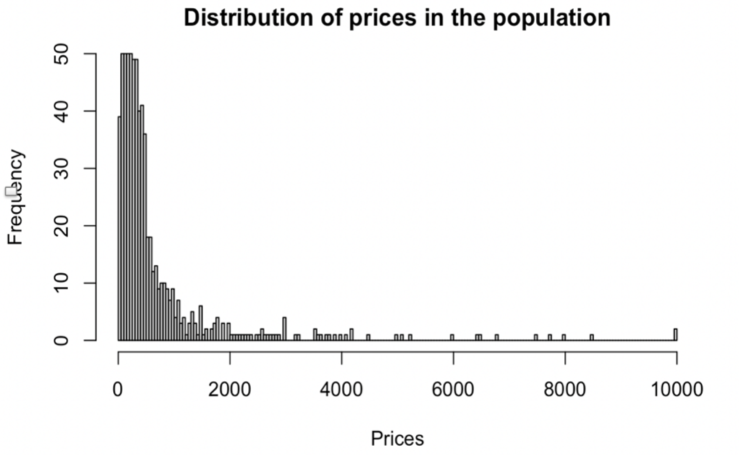


Figure3. Distribution of Airbnb's prices in population

The three figures above show the distribution of Airbnb's prices' distribution under SRS sampling, STR sampling and true population. We find that the stratified sampling has a distribution more like the actual distribution compared to that of simple random sampling. Therefore, the stratified sample can represent the price characteristics better than the simple random sample.

However, the stratified sample does not contain extreme cases (e.g., the case when price=10000). Due to the small number of extremity values in the population, the current sample size (n=1200) does not include the case when price=10000, and the problem cannot be easily solved by simply increasing the sample size.

We also investigated the mean prices of each stratum and calculated the respective counts of Airbnb in each stratum. Here, we take neighbourhood groups as strat and plot the distributions of the average price of the five strata obtained from the two sampling methods. The three graphs below are the population distribution of the average mean price of each stratum, the distribution of simple random sampling of the average mean price of each stratum and the distribution result using the stratified sampling method, respectively. The distribution of mean prices of Airbnb of the stratified sample in Queens, Brooklyn, and Manhattan are very similar to those of the population since we have larger counts of Airbnb in these regions and the resulting stratum means are more representative.

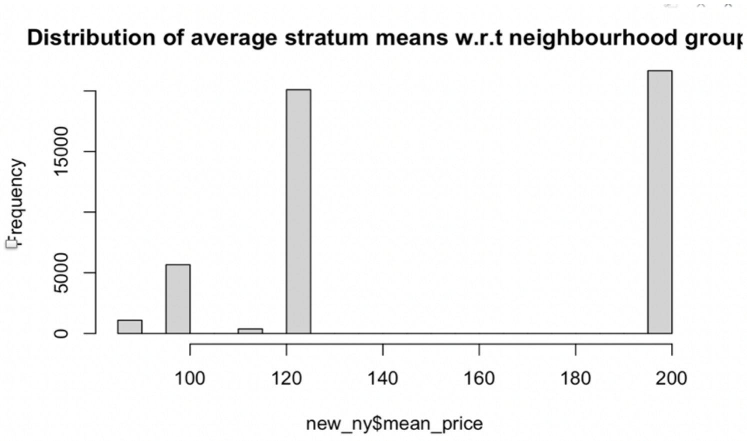


Figure 4. Average Airbnb's prices and Number of Airbnb's price in the 5 neighborhood groups respectively

From left to right: 'Bronx','Queens','Staten Island','Brooklyn','Manhattan'

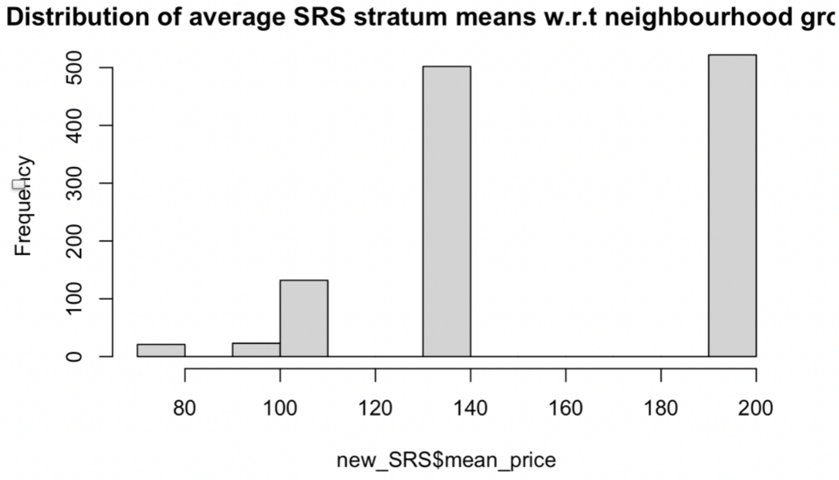


Figure 5. Average Airbnb's prices and Number of Airbnb's price in the 5 neighborhood groups get from SRS respectively

From left to right: 'Staten Island','Bronx','Queens',''Brooklyn','Manhattan'

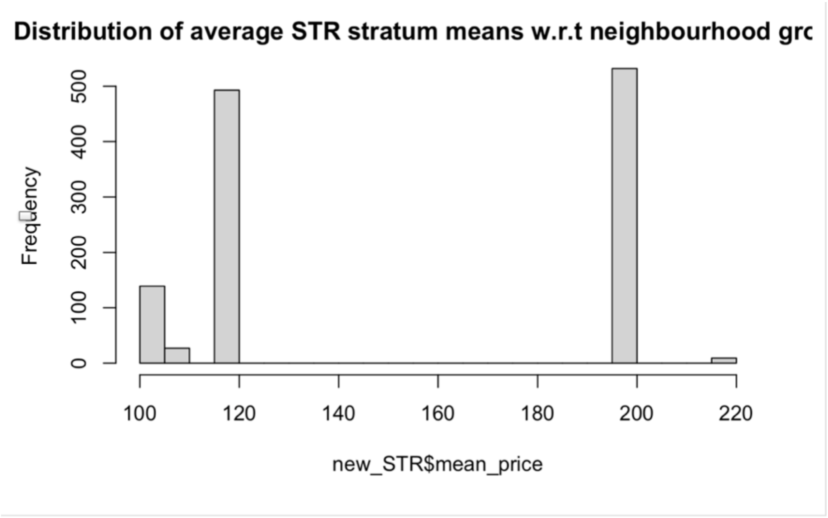


Figure 6. Average Airbnb's prices and Number of Airbnb's price in the 5 neighborhood groups get from STR respectively

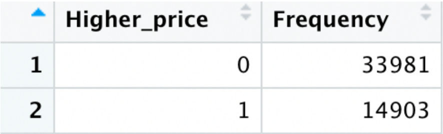
From left to right: 'Queens','Bronx','Brooklyn','Manhattan','Staten Island'

After looking at the mean price of Airbnb in New York, we take a closer look at the proportion of Airbnb that are more expensive than the mean price of \$151.26. We use the mean price of stratified sampling since it's considered to be more representative.

Prices higher than or the same as the average price are considered expensive and are assigned by value "1", and prices lower than the average price are denoted by "0". From the simple random sample, we use the equations: $E(\hat{p}_S) = p_P$, SE($\hat{p}_S$)=$\sqrt{\frac{(1-\hat{p}_S)\hat{p}_S}{n}}$, $95\%\ CI = \hat{p}_S \pm 1.96 * \text{SE}(\hat{p}_S)$, where $\hat{p}_S$ is the sample proportion of prices higher than the sample mean of the stratified sample. We found that around 32% of the sampled Airbnb has a higher price than \$151.26, with a standard error of about 1.33%. On the other hand, by observing the data result from stratified sampling, we use the equations: $\widehat{p_{str}} =$

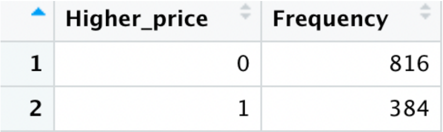
$\sum_{h=1}^{5}\frac{N_h}{N}\widehat{p_{sh}}$, SE($\widehat{p_{str}}$)= $\sqrt{\sum_{h=1}^{5}\frac{N_h^2}{N^2}(1-\frac{n_h}{N_h})\frac{\widehat{p_{sh}}(1-\widehat{p_{sh}})}{n_h}}$, $95\%\ CI = \widehat{p_{str}} \pm 1.96 * \mathrm{SE}(\widehat{p_{str}})$, where , $\widehat{p_{str}}$ is the sample proportion of higher prices of all stratum, $\widehat{p_{sh}}$ is the sample proportion of higher prices of each strata. We find that about 32.58% of Airbnb in the sample has higher prices than the average price, which we consider expensive. The stratified sampling standard error of proportional of expensive Airbnb is around 1.29%. Since the standard error of stratified sampling is smaller, we believe it is more accurate. The 95% confidence interval for a simple random sample is [29.39%,34.61%]. Among 100 times of repeated sampling, there are approximately 95 times that the true population proportion of expensive Airbnb in New York lies between 29.39% and 34.61%. The confidence interval computed for the stratified sampling method is [30.01%,35.10%]. In other words, in repeated samples, we capture the actual value of the population of expensive Airbnb in New York between 30.01% and 35.10%, about 19 out of 20 times.

In reality, the true population information is often not known at advance, we would say stratified sampling is better because of the smaller standard error. However, in the current case, we notice that the value we get from using simple random sampling is closer to the population proportion. Thus, to compare the fitness of the two sampling methods, we take a closer look at the ratio between the number of observations that are considered "cheap" to "expensive."



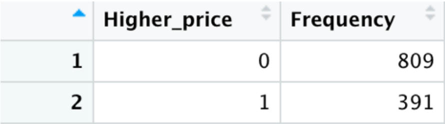
	Higher_price	Frequency
1	0	33981
2	1	14903

table1: population counts of binary variables



	Higher_price	Frequency
1	0	816
2	1	384

table2: counts of binary variables from SRS



	Higher_price	Frequency
1	0	809
2	1	391

table3: counts of binary variables from STR

The tables above depict the respective counts for binary data received from the population, simple random sampling, and stratified sampling. Comparing the ratio of counts of higher price to counts of lower price reveals that the ratio obtained from SRS 384/816 is closer to the ratio of the population, which is 114903/33981.

Comparing two distinct sampling approaches, if population information is unknown in advance, it is clear that stratified sampling works better with both parameters than simple random sampling for our study due to smaller standard errors. Second, based on the stratified sample calculations, we estimate that the average Airbnb price is $151.26, and around 32.58 percent of them are deemed expensive. Thus, it provides some reference values for the pricing of average hotels in New York, as well as some lodging prices for travellers who wish to visit New York, enabling them to allocate their trip budgets rationally.

However, there are some limitations to our study. Firstly, since the population size of Staten Island is small in the dataset, which is only 373, the conclusion may contain some bias due to a lack of information. Secondly, this data is selected for only one year and is not the most up to date. Also, statistically speaking, in the binary case there can be a limitation result from proportional allocation. Proportional allocation stratified sampling can ensure that $\frac{n_h}{N_H} = \frac{n}{N}$, but there may be some biased cases. For instance, there are neighborhood groups with larger counts of Airbnb prices and they also have higher average stratum prices (for example, Manhattan and Brooklyn). When these neighborhood groups occupy a large portion of the sample size, it can result in an overestimation of the proportion of higher Airbnb prices. In other words, proportional allocation cannot guarantee that the ratio of 1 to 0 of the sample higher price proportion is a good representation of the population ratio. Finally, proportional allocation does not provide a particular real-time dimension and does not consider the impact of factors such as inflation on prices, which may lead to certain bias and errors. Moreover, the dataset only includes Airbnbs in New York. New York, being the financial center of the United States and a popular tourist destination, cannot speak for other cities in the country. In addition, the average price level and living expenses in New York are considerably greater than in other locations, such as the central United States. Consequently, the following result for Airbnb costs in New York may not be the most accurate depiction of Airbnb prices across the nation. Therefore, the conclusion cannot be extrapolated to the United States in this instance.

PART II

On the field of multi-parameter hypothesis testing, many researchers have argued that the likelihood ratio test (LRT) produces some bias and is somewhat flawed because the size alpha LRT is not similar on the boundary of the null hypothesis. Thus, they offer a variety of new tests and claim that all these tests are superior to the LRT. These new tests acknowledge the problems with the Neyman-Pearson theory of impartiality and the most powerful size alpha test and alpha-admissibility criteria. However, these criteria violate statistical common sense, are statistically unacceptable and have no practical use. Perlman and Wu s' paper demonstrates the problems and flaws of the new tests and justifies the LRT. Additionally, they acknowledge that the LR criterion is not perfect, but it is still the preferred method in the case of non-Bayesian parametric hypothesis testing. (Perlman, 1999)

APPENDIX

R Code:

Dataset

~~~
```{r}
library(arsenal)
ny<-read.csv('/Users/bessie/Desktop/AB_NYC_2019.csv')
ny=filter(test,price!=0)
ny=transform(ny,higher_price=ifelse(price>=152.7551,1,0))
```
~~~

Simple random sampling and stratified sampling with respect to neighborhood groups (continuous variable: price) :

~~~
set.seed(20)
attach(ny)
price<-ny$price
neighborhood_group<-ny$neighbourhood_group
N.h <- tapply(price,neighborhood_group,length)
##N.h
#population size for different regions
 neighborhood_group<- names(N.h) # name of the regions
##eduyrs
N<-sum(N.h)
##N
n<-1200
##simple random sample
set.seed(20)
SRS.indices <- sample.int(N,   n, replace = F)
SRS.sample <- ny[SRS.indices , ]
ybar.srs <- mean(SRS.sample$price)
se.srs <- sqrt((1 - n / N) * var(SRS.sample$price) / n)
srs <- c(ybar.srs, se.srs)
srs
##Stratified sampling using proportional allocation
n.h.prop <- round( (N.h/N) * n)
STR.sample.prop <- NULL
set.seed(20)
for (i in 1: length(neighborhood_group))
{
  row.indices <- which(ny$neighbourhood_group == neighborhood_group[i])
  sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
  STR.sample.prop <- rbind(STR.sample.prop, ny[sample.indices, ])
}

ybar.h.prop <- tapply(STR.sample.prop$price, STR.sample.prop$neighbourhood_group, mean)
var.h.prop <- tapply(STR.sample.prop$price, STR.sample.prop$neighbourhood_group, var)
##length of N.h=16, but length of var.h.prop & average income=14
se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
##rbind(ybar.h.prop, se.h.prop)
ybar.str.prop <- sum(N.h / N * ybar.h.prop)
se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
str.prop <- c(ybar.str.prop, se.str.prop)
str.prop
```
~~~

Simple random sampling and stratified sampling with respect to neighborhood groups (Binary variable: price>=stratified sample mean or not) :

~~~
```{r}
set.seed(20)
ny=transform(ny,higher_price=ifelse(price>=151.257781,1,0))
attach(ny)
higher_price<-ny$higher_price
neighborhood_group<-ny$neighbourhood_group
N.h <- tapply(higher_price,neighborhood_group,length)
##N.h
#population size for different regions
neighborhood_group<- names(N.h) # name of the regions
##eduyrs
N<-sum(N.h)
##N
n<-1200
##simple random sample
set.seed(20)
SRS.indices <- sample.int(N,   n, replace = F)
SRS.sample <- ny[SRS.indices , ]
ybar.srs <- mean(SRS.sample$higher_price)
se.srs <- sqrt((1 - n / N) * var(SRS.sample$higher_price) / n)
srs <- c(ybar.srs, se.srs)
srs
c(ybar.srs-1.96*se.srs,ybar.srs+1.96*se.srs)
##Stratified sampling using proportional allocation
n.h.prop <- round( (N.h/N) * n)
~~~

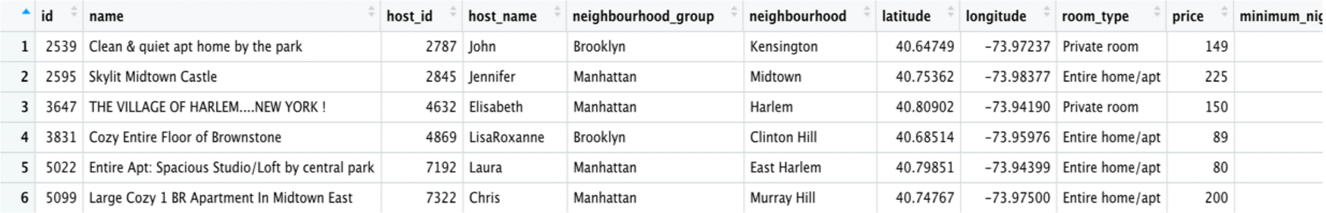
~~~
set.seed(20)
for (i in 1: length(neighborhood_group))
{
  row.indices <- which(ny$neighbourhood_group == neighborhood_group[i])
  sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
  STR.sample.prop <- rbind(STR.sample.prop, ny[sample.indices, ])
}
ybar.h.prop <- tapply(STR.sample.prop$higher_price, STR.sample.prop$neighbourhood_group, mean)
var.h.prop <- tapply(STR.sample.prop$higher_price, STR.sample.prop$neighbourhood_group, var)
se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
##rbind(ybar.h.prop, se.h.prop)
ybar.str.prop <- sum(N.h / N * ybar.h.prop)
se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
str.prop <- c(ybar.str.prop, se.str.prop)
str.prop
c(ybar.str.prop-1.96*se.str.prop,ybar.str.prop+1.96*se.str.prop)
##Population average:
mean(ny$higher_price)
```
Graphical illustration of prices (Population, SRS sample & Stratified sample):
```{r}
## Population distribution of prices
ppl_counts=as.data.frame(table(ny$price))
ppl_counts$Var1 <- as.numeric(as.character(ppl_counts$Var1))
names(ppl_counts)=c('Price','Frequency')
hist(ppl_counts$Price,breaks=200,main='Distribution of prices in the population',xlab='Prices',ylab='Frequency')
## SRS sample distribution of prices
sample_counts=as.data.frame(table(SRS.sample$price))
sample_counts$Var1 <- as.numeric(as.character(sample_counts$Var1))
names(sample_counts)=c('Price','Frequency')
hist(sample_counts$Price,breaks=200,main='Distribution of prices in the SRS sample',xlab='Prices',ylab='Frequency')
## Stratified sample distribution of prices
str_sample_counts=as.data.frame(table(STR.sample.prop$price))
str_sample_counts$Var1 <- as.numeric(as.character(str_sample_counts$Var1))
names(str_sample_counts)=c('Price','Frequency')
hist(str_sample_counts$higher_price,breaks=200,main='Distribution of prices in the stratified
sample',xlab='Prices',ylab='Frequency')
```
Graphical illustration of mean prices (Population)
```{r}
## Population distribution of mean prices over different neighborhood groups.
ppl_neigh_counts=as.data.frame(table(ny$neighbourhood_group))
names(ppl_neigh_counts)=c('neighbourhood_group','frequencies')
ppl_meanP <- ny %>%
  group_by(neighbourhood_group) %>%
  summarize(mean_price=mean(price))
ppl_meanP=merge(ppl_neigh_counts,ppl_meanP,by='neighbourhood_group')
new_ny=merge(ppl_meanP,ny,by='neighbourhood_group')
h1=hist(new_ny$mean_price,main='Distribution of average stratum means w.r.t neighbourhood groups')
text(h1$mids,h1$neighbourhood_group,labels=TRUE)
```
Graphical illustration of mean prices (SRS sample)
```{r}
## STR distribution of mean prices over different neighborhood groups.
SRS_neigh_counts=as.data.frame(table(SRS.sample$neighbourhood_group))
names(SRS_neigh_counts)=c('neighbourhood_group','frequencies')
SRS_meanP <- SRS.sample %>%
  group_by(neighbourhood_group) %>%
  summarize(mean_price=mean(price))
SRS_meanP=merge(SRS_neigh_counts,SRS_meanP,by='neighbourhood_group')
new_SRS=merge(SRS_meanP,SRS.sample,by='neighbourhood_group')
h2=hist(new_SRS$mean_price,main='Distribution of average SRS stratum means w.r.t neighbourhood groups')
text(h2$mids,h2$neighbourhood_group,labels=TRUE)
```
Graphical illustration of mean prices (Stratified sample)
```{r}
## STR distribution of mean prices over different neighborhood groups.
STR_neigh_counts=as.data.frame(table(STR.sample.prop$neighbourhood_group))
~~~

~~~~
STR_meanP <- STR.sample.prop %>%
  group_by(neighbourhood_group) %>%
  summarize(mean_price=mean(price))
STR_meanP=merge(STR_neigh_counts,STR_meanP,by='neighbourhood_group')
new_STR=merge(STR_meanP,STR.sample.prop,by='neighbourhood_group')
h3=hist(new_STR$mean_price,main='Distribution of average STR stratum means w.r.t neighbourhood groups',breaks=20)
text(h3$mids,h3$neighbourhood_group,labels=TRUE)
```
~~~~

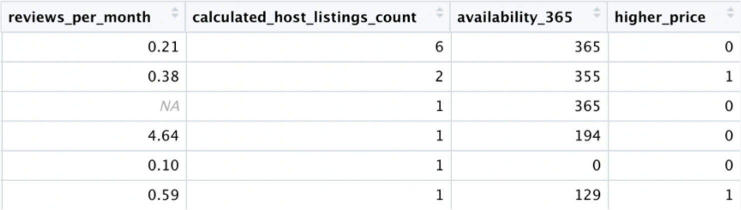
Table illustration of the binary variable of prices (Population, SRS sample & Stratified sample):

~~~~
```{r}
## Population distribution of prices
ppl_bi_counts=as.data.frame(table(ny$higher_price))
ppl_bi_counts$Var1 <- as.numeric(as.character(ppl_bi_counts$Var1))
names(ppl_bi_counts)=c('Higher_price','Frequency')
View(ppl_bi_counts)
## SRS sample distribution of prices
sample_bi_counts=as.data.frame(table(SRS.sample$higher_price))
sample_bi_counts$Var1 <- as.numeric(as.character(sample_bi_counts$Var1))
names(sample_bi_counts)=c('Higher_price','Frequency')
View(sample_bi_counts)
## Stratified sample distribution of prices
str_sample_bi_counts=as.data.frame(table(STR.sample.prop$higher_price))
str_sample_bi_counts$Var1 <- as.numeric(as.character(str_sample_bi_counts$Var1))
names(str_sample_bi_counts)=c('Higher_price','Frequency')
View(str_sample_bi_counts)
~~~~

Dataset Overview



▲	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nig
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	
2	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	
3	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	
4	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	
5	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	
6	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	



reviews_per_month	calculated_host_listings_count	availability_365	higher_price
0.21	6	365	0
0.38	2	355	1
NA	1	365	0
4.64	1	194	0
0.10	1	0	0
0.59	1	129	1

Citation:

Dataset:

Dgomonov. (2019, August 12). *New York City airbnb open data*. Kaggle. Retrieved November7, 2022, from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?select=New_York_City_.png&page=2

Dgomonov. (2020, August 3). *Data Exploration on NYC airbnb*. Kaggle. Retrieved November 11, 2022, from https://www.kaggle.com/code/dgomonov/data-exploration-on-nyc-airbnb

Reading:

Perlman, M. D., & Wu, L. (1999). The emperor's new tests. *Statistical Science, 14*(4), 355-369. https://doi.org/10.1214/ss/1009212517

Perlman, M. D., & Wu, L. (1999). The emperor's new tests. *Statistical Science, 14*(4), 355369. https://doi.org/10.1214/ss/1009212517