The University of British Columbia

STAT 344 Project

Yuhao Zhou	12265294
Cong Cheng	31070519
Yandong Huang	73422875
Qin(Jean) Xue	72070014

Leader: Yuhao Zhou

Contribution:

We decided the domain of the project topic, and sampling methods together.
Yuhao Zhou wrote the code and report.
Cong Cheng wrote the code and report.
Yandong Huang wrote the code and report.
Qin Xue wrote the Part II and report.

Abstract

The average price of used cars in Poland are collected for the new driver who wants to practise their driving skills on the used car and for those who cannot afford a new car. To make an analysis, simple random sampling and stratified sampling are applied. The transmission type and price are used to estimate the average price of the used car. By comparing standard error and confidence interval, the stratified sample probably performs the best. Also for the stratified sample, transmission is the feature which is applied for stratification and being tested with different proportions for two types of transmission.

Introduction

Nowadays transportation has become an inevitable part for human beings, the cars are the main transportation for most of the families. Especially for university students, many of them just get their driver's licence, and do not have enough confidence to drive a new car with huge cost. The used car will become their best choice for transition to a new car in the future. At this time, the average used car price will provide them an idea about how much the car is worth, and whether they got deceived by the car seller. Also for those who wish to drive the luxury car, but cannot afford the price for the new arrival, the second-hand car will become a perfect choice for them. The project also analyses the proportion of the luxury car in the Used Car market for those people who have a concept about whether the car is luxury or not.

In this project, a dataset "Used Car Price Prediction" downloaded from Kaggle (https://www.kaggle.com/code/iabhishekmaurya/used-car-price-prediction/data?select=train-data.csv) (ABHISHEK, 2020) are analysed for the best sampling methods. It includes all used cars that were sold in Poland in 2020. The dataset involves 14 variables and 6019 observations in total. The variable of interest is Price, which is a continuous variable. Besides that there is another variable of interest is whether the price is lower than 12.999 INR Lakhs as a discrete variable, which is also defined as is this used car a luxury car or not. In order to determine the average price for all cars sold in Poland, price is treated as the response variable and other variables are potential factors.

Among 6019 observations, there is 1 observation involving unreasonable value. Therefore, that observation was omitted and the analysis of the best sampling method would be down on the remaining 6018 observations.

1.1 Variable selection and descriptions

In the dataset, there are 14 variables in total. The price of the used car in INR Lakhs, ranging from 0 to 160, is a continuous variable and would be the response variable to be predicted in this project. Within the other 13 variables, we choose transmission as the explanatory variable to help us make a stratified sample. This explanatory variable is listed and described in the following table.

Table 1. Explanatory Variable description

Variable	Description
transmission	The type of transmission(automatic and manual)

price	The price of used car

1.2 Variable summary statistics

Table 2 proportion in transmission

	Automatic	Manual
Frequency	1719	4299
Proportion	0.286	0.714

Table 2 shows the proportion for each type of transmission.

The summary statistics for each continuous variable is shown in the following table.
Table 3. Summary statistics for all continuous variables after transformation

Variable	# of obs	Mean	Median	SD	Min	Max
Price	6018	9.47	5.64	11.17	0.44	160

2. Methods

Two sampling methods, simple random sampling and stratified sampling, were used to draw a sample from the population and use the sample mean to estimate the population mean. After applying different sampling methods, standard error and confidence interval will be calculated and compared. The method with the best results in terms of standard error and confidence interval would be the method that fits the Used Car Data best.

2.1 Simple Random Sample

A simple random sample is a sampling method which chooses a subset randomly and each observation with equal possibility. First randomly sampled 300 observations from the population, as the population size is over 6000, therefore, FPC is ignored in this situation. For the estimation in a simple random sample, vanilla estimation was used to get standard error and confidence interval. Here is the formula used to calculate the standard error and confidence interval for the continuous variable in Vanilla estimation.

$SE(y_s) = \sqrt{\frac{S_s^2}{n}}$ $\quad CI(y_s) = \overline{y_s} \pm 1.96\, SE(y_s)$

Here is the formula used to calculate the standard error and confidence interval for the discrete variable in Vanilla estimation. In other words, we are calculating the percentage of cars which have a price lower than 12.9999.

$SE(p_s) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\quad CI(p_s) = \overline{p_s} \pm 1.96SE(p_s)$

The coding is in Appendix (see Appendix Figure 1).

2.2 Stratified Sample

Stratified random sampling is a method of sampling that involves the division of a population into smaller subgroups known as strata. There are two approaches for the stratified sampling method, the first applied is stratify with equal proportion. The second is the proportion allocation approach for stratified samples.

2.2.1 Equal proportion

First draw 300 observations from the population, which is 150 from both manual and automatic. The standard error and confidence interval for the continuous variable for each stratum is calculated to estimate the sample standard error and confidence interval with the following formula.

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{s_{s_h}^2}{n_h}} \quad CI(\bar{y}_{str}) = \bar{y}_{str} \pm 1.96SE(\bar{y}_{str})$$

The standard error and confidence interval for the discrete variable for each stratum is calculated to estimate the sample standard error and confidence interval with the following formula.

$$SE(\bar{p}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}} \quad CI(\bar{p}_{str}) = \bar{p}_{str} \pm 1.96SE(\bar{p}_{str})$$

The coding is in Appendix (see Appendix Figure 2).

2.2.2 Proportion allocation

The main difference between two types of stratified sample is the proportion for each stratum. In this situation, 300 observations are sampled from the population, which is 90 from automatic, and 210 from manual. As it is shown in 1.3, the proportion between manual and automatic is approximately 7:3. The standard error and confidence interval for the continuous variable for each stratum is calculated to estimate the sample standard error and confidence interval with the following formula.

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{s_{s_h}^2}{n_h}} \quad CI(\bar{y}_{str}) = \bar{y}_{str} \pm 1.96SE(\bar{y}_{str})$$

The standard error and confidence interval for the discrete variable for each stratum is calculated to estimate the sample standard error and confidence interval with the following formula.

$$SE(\bar{p}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}} \quad CI(\bar{p}_{str}) = \bar{p}_{str} \pm 1.96SE(\bar{p}_{str})$$

The coding is in Appendix (see Appendix Figure 3).

3. Results

In the project, standard error and confidence intervals are used to compare among different sampling methods. As for the confidence intervals, we applied 95% confidence intervals in the project. Standard error is a statistic that is the approximate standard deviation of a statistical sample population.

Table 4 95% confidence interval, estimation, standard error for all three sampling methods for the continuous variable

method	lower	estimation	upper	SE
Simple Random Sample	9.018934	10.3751667	11.73140	0.6919557
Stratified with equal proportion	8.864489	9.7718667	10.67924	0.4629478
Stratified with proportion allocation	8.386401	9.5275984	10.66880	0.5822438

As can be seen from the table above, the stratified sample with equal proportion has the smallest SE and narrow confidence interval. Simple random sample has the worst performance based on the table shown. All of them have a 95% confidence interval near 9-11.

Table 5 95% confidence interval, estimation, standard error for all three sampling methods for the discrete variable

method	lower	estimation	upper	SE
Simple Random Sample	0.7331236	0.78	0.8268764	0.02391652
Stratified with equal proportion	0.7642327	0.79718179	0.8301308	0.01681075
Stratified with proportion allocation	0.7737667	0.81023833	0.8467100	0.01860798

As can be seen from the table above, the stratified sample with equal proportion has the smallest SE and narrow confidence interval. Simple random sample has the worst performance based on the table shown. The result is quite similar to the one with the continuous variable. Also all of them have an estimation for the proportion for car prices lower than 12.9999 is around 80%.

4. Discussion

According to table 4 and table 5, stratified samples with both equal proportion and proportion allocation perform well. Since both stratified samples have a narrower confidence interval and smaller standard error. Within two stratified samples, equal proportion has a smaller standard error than the proportion allocation, it can be concluded that the stratified with equal proportion fits the Used Car Price data best.

4.1 Simple random sample

As can be seen from table 4 and 5, simple random sample method always performs the worst. That is because the simple random samples are sampling observations randomly from the population, but in the population the proportion for automatic and manual cars are not equal, and automatic cars share a higher car price than manual. In a simple random sample, the estimation of sample mean didn't consider the proportion for each transmission, so the performance for this sampling method is terrible.

4.2 Stratified with equal proportion

As shown in Table 4 and 5, stratified with equal proportion has the best performance among all three sampling methods. The division for two types of transmission plays an important role in that performance. The contribution of the sample is 50% with manual and 50% with automatic. Although the price is different for each transmission, the proportion is equal which represents that sample will not be influenced by a specific type. That is the reason for a great performance in a Stratified sample with equal proportion.

4.3 Stratified with proportion allocation

For the results above, stratified with proportion allocation didn't perform as great as the assumption. The standard error is slightly higher than the one with equal proportion in both continuous and discrete variables. The sample was chosen by proportion, therefore, 30% are automatic and 70% are manual. There is an interesting phenomenon in the data set for Used Car Price. The distribution of price for the main population is extremely skewed, and all of the extreme values are way higher than the average of the price in the population. For all of the extreme values, their transmission type is automatic. It's also the reason that estimation in proportion allocation is higher than the equal proportion. That explains why the proportion allocation sampling method performs worse than the equal proportion. The existence of extremes affect the standard error a lot.

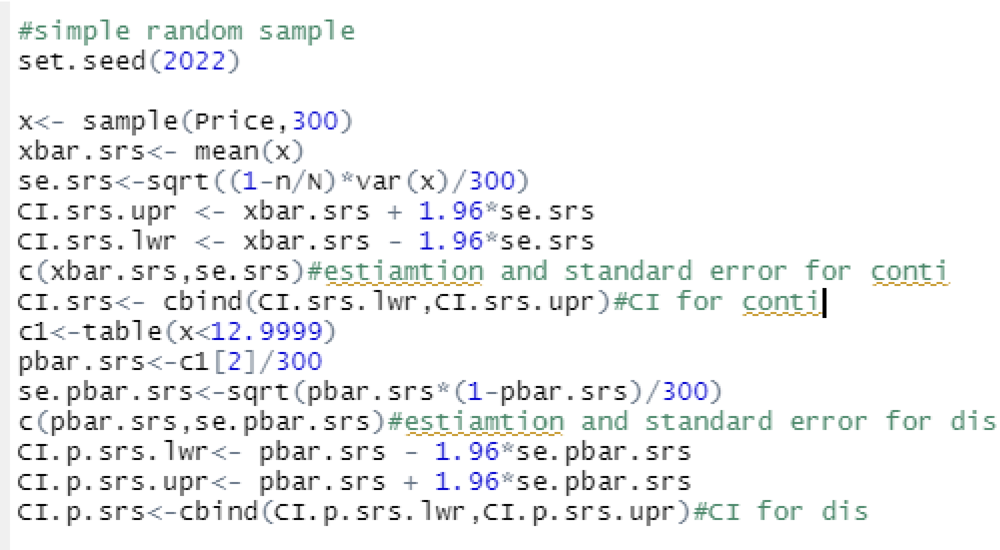
5. Conclusion

Based on the two criteria (standard error, 95% confidence interval) which measures those sampling methods, the stratified sample with equal proportion is the most suitable sampling method for the Used Car Price dataset on estimating the population mean, the average of price for used car. With the estimation on average price 9.77 INR Lakhs. This sampling method is also the most suitable method for the discrete variable, whether the price of a used car is lower than 12.9999 INR Lakhs. The probability that a used car is lower than 12.9999 INR Lakhs is 79.72%.

The results could be generalized to a larger population, but there exists some limitations. For the simple random sample, if only sampled once, the result may not be representative for a dataset that has two obvious stratum with huge difference in proportion. If the variance of interest of variable is relatively small, proportion allocation may perform better than equal proportion.

References

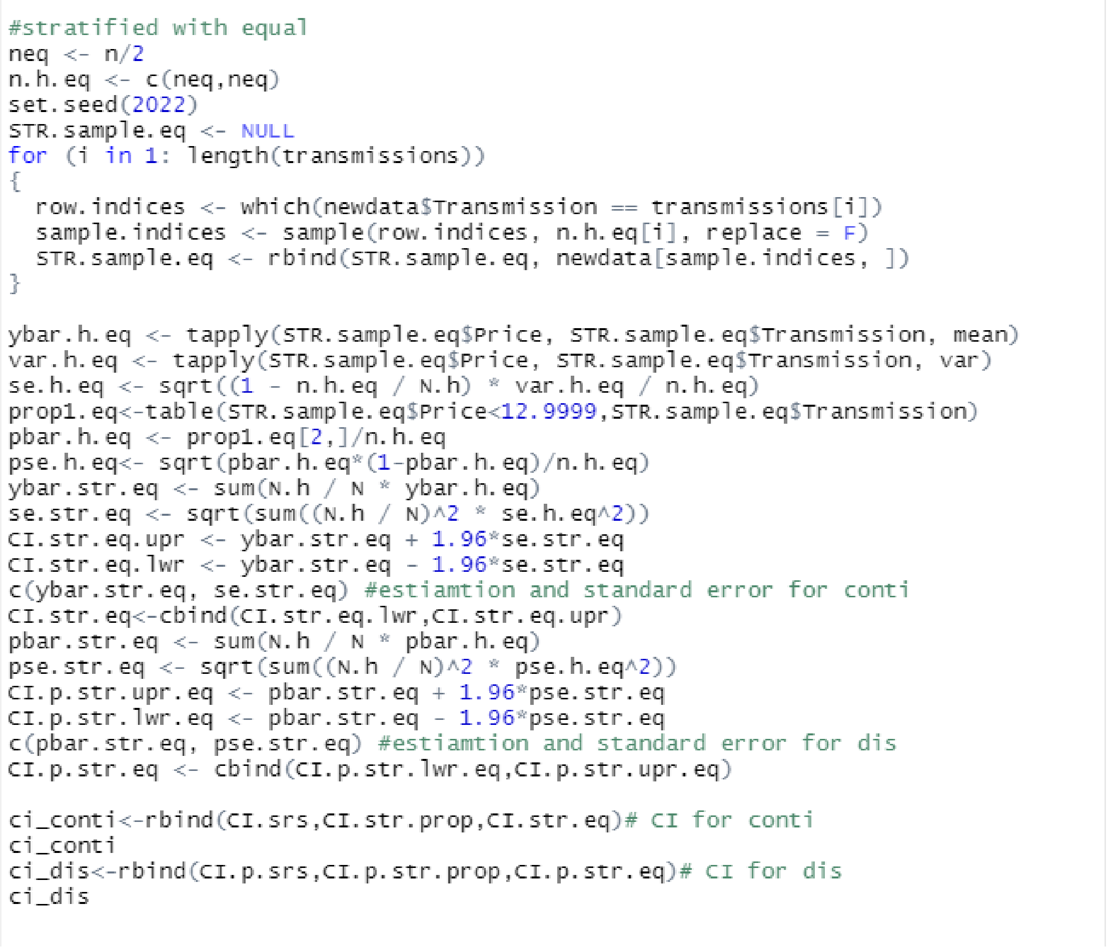
Abhishek (2020). *Used Car Price Prediction*. kaggle. https://www.kaggle.com/code/iabhishekmaurya/used-car-price-prediction/data?select=train-data.csv



```
#simple random sample
set.seed(2022)

x<- sample(Price,300)
xbar.srs<- mean(x)
se.srs<-sqrt((1-n/N)*var(x)/300)
CI.srs.upr <- xbar.srs + 1.96*se.srs
CI.srs.lwr <- xbar.srs - 1.96*se.srs
c(xbar.srs,se.srs)#estiamtion and standard error for conti
CI.srs<- cbind(CI.srs.lwr,CI.srs.upr)#CI for conti
c1<-table(x<12.9999)
pbar.srs<-c1[2]/300
se.pbar.srs<-sqrt(pbar.srs*(1-pbar.srs)/300)
c(pbar.srs,se.pbar.srs)#estiamtion and standard error for dis
CI.p.srs.lwr<- pbar.srs - 1.96*se.pbar.srs
CI.p.srs.upr<- pbar.srs + 1.96*se.pbar.srs
CI.p.srs<-cbind(CI.p.srs.lwr,CI.p.srs.upr)#CI for dis
```

(Figure 1 code for simple random sample)

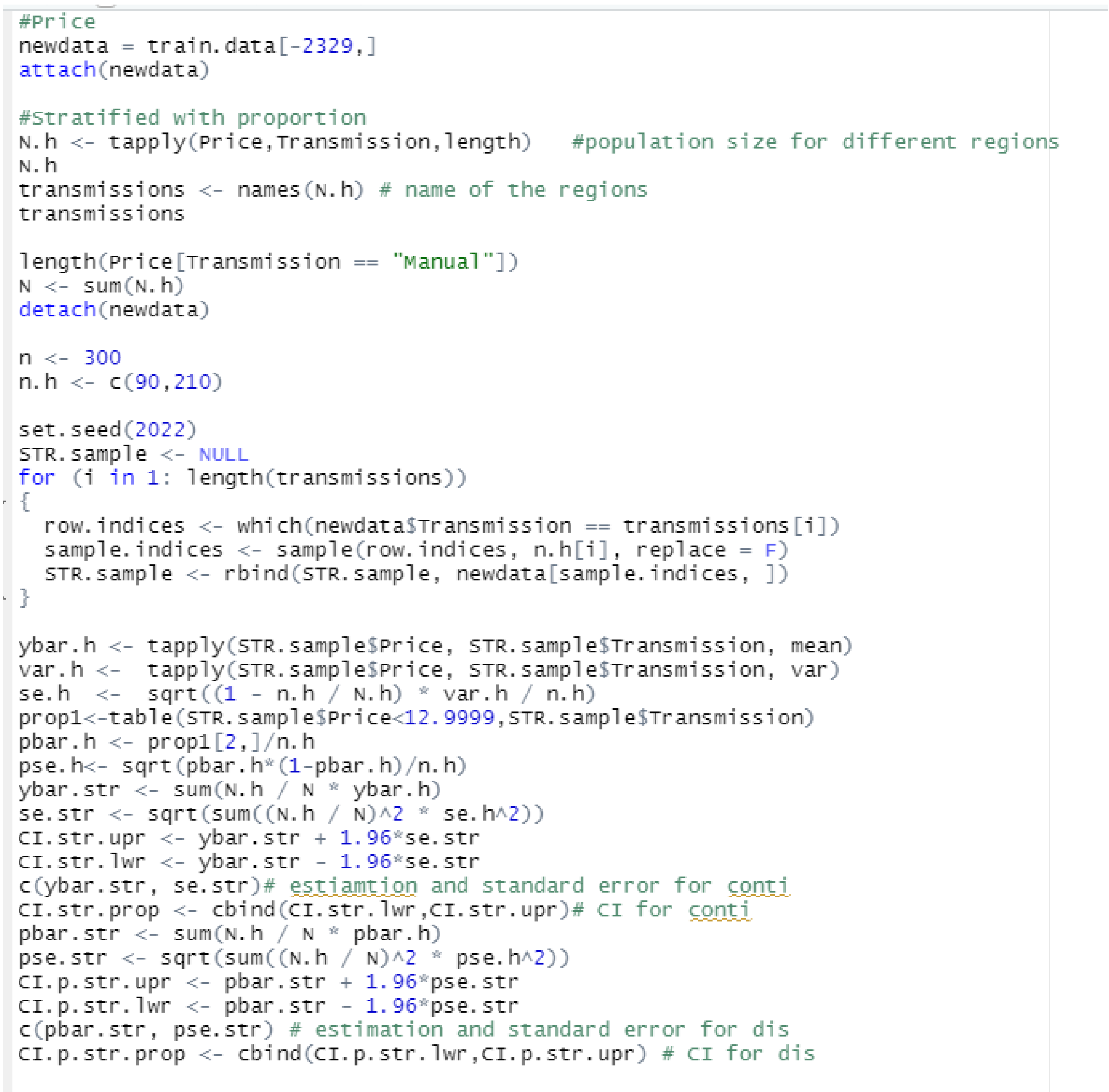


```
#stratified with equal
neq <- n/2
n.h.eq <- c(neq,neq)
set.seed(2022)
STR.sample.eq <- NULL
for (i in 1: length(transmissions))
{
  row.indices <- which(newdata$Transmission == transmissions[i])
  sample.indices <- sample(row.indices, n.h.eq[i], replace = F)
  STR.sample.eq <- rbind(STR.sample.eq, newdata[sample.indices, ])
}

ybar.h.eq <- tapply(STR.sample.eq$Price, STR.sample.eq$Transmission, mean)
var.h.eq <- tapply(STR.sample.eq$Price, STR.sample.eq$Transmission, var)
se.h.eq <- sqrt((1 - n.h.eq / N.h) * var.h.eq / n.h.eq)
prop1.eq<-table(STR.sample.eq$Price<12.9999,STR.sample.eq$Transmission)
pbar.h.eq <- prop1.eq[2,]/n.h.eq
pse.h.eq<- sqrt(pbar.h.eq*(1-pbar.h.eq)/n.h.eq)
ybar.str.eq <- sum(N.h / N * ybar.h.eq)
se.str.eq <- sqrt(sum((N.h / N)^2 * se.h.eq^2))
CI.str.eq.upr <- ybar.str.eq + 1.96*se.str.eq
CI.str.eq.lwr <- ybar.str.eq - 1.96*se.str.eq
c(ybar.str.eq, se.str.eq) #estiamtion and standard error for conti
CI.str.eq<-cbind(CI.str.eq.lwr,CI.str.eq.upr)
pbar.str.eq <- sum(N.h / N * pbar.h.eq)
pse.str.eq <- sqrt(sum((N.h / N)^2 * pse.h.eq^2))
CI.p.str.upr.eq <- pbar.str.eq + 1.96*pse.str.eq
CI.p.str.lwr.eq <- pbar.str.eq - 1.96*pse.str.eq
c(pbar.str.eq, pse.str.eq) #estiamtion and standard error for dis
CI.p.str.eq <- cbind(CI.p.str.lwr.eq,CI.p.str.upr.eq)

ci_conti<-rbind(CI.srs,CI.str.prop,CI.str.eq)# CI for conti
ci_conti
ci_dis<-rbind(CI.p.srs,CI.p.str.prop,CI.p.str.eq)# CI for dis
ci_dis
```

(Figure 2 code for Stratified sample with equal proportion)



```
#Price
newdata = train.data[-2329,]
attach(newdata)

#Stratified with proportion
N.h <- tapply(Price,Transmission,length)   #population size for different regions
N.h
transmissions <- names(N.h) # name of the regions
transmissions

length(Price[Transmission == "Manual"])
N <- sum(N.h)
detach(newdata)

n <- 300
n.h <- c(90,210)

set.seed(2022)
STR.sample <- NULL
for (i in 1: length(transmissions))
{
  row.indices <- which(newdata$Transmission == transmissions[i])
  sample.indices <- sample(row.indices, n.h[i], replace = F)
  STR.sample <- rbind(STR.sample, newdata[sample.indices, ])
}

ybar.h <- tapply(STR.sample$Price, STR.sample$Transmission, mean)
var.h <-  tapply(STR.sample$Price, STR.sample$Transmission, var)
se.h  <-  sqrt((1 - n.h / N.h) * var.h / n.h)
prop1<-table(STR.sample$Price<12.9999,STR.sample$Transmission)
pbar.h <- prop1[2,]/n.h
pse.h<- sqrt(pbar.h*(1-pbar.h)/n.h)
ybar.str <- sum(N.h / N * ybar.h)
se.str <- sqrt(sum((N.h / N)^2 * se.h^2))
CI.str.upr <- ybar.str + 1.96*se.str
CI.str.lwr <- ybar.str - 1.96*se.str
c(ybar.str, se.str)# estiamtion and standard error for conti
CI.str.prop <- cbind(CI.str.lwr,CI.str.upr)# CI for conti
pbar.str <- sum(N.h / N * pbar.h)
pse.str <- sqrt(sum((N.h / N)^2 * pse.h^2))
CI.p.str.upr <- pbar.str + 1.96*pse.str
CI.p.str.lwr <- pbar.str - 1.96*pse.str
c(pbar.str, pse.str) # estimation and standard error for dis
CI.p.str.prop <- cbind(CI.p.str.lwr,CI.p.str.upr) # CI for dis
```

(Figure 3 coding for stratified sample with proportion allocation)

Part II

There is an empire. The emperor is especially fond of the likelihood ratio test (LRT) among a bunch of statistical inference tools to find the truth and reach a reasonable conclusion to help him make a better decision for problems. However, a bright young man noticed the LRT would not infer accurate conclusions for some particular problems, so he constructed a New Test which could infer a more accurate conclusion for those particular problems. Then the New Test replaced LRT. But one day, the emperor found the New Test violated statistical intuition and not suitable for most problems. Then the emperor abandoned the New Test and restored the LRT. This story is not completely fictional. Going back to reality, there are many “New Tests” arised and seem better than LRT. Marden, Perlman and Stein show that LRT is not only d-inadmissible, but may also be “worse than useless” on some particular problems. Showing the LRT is not universally satisfactory for all problems. But those “New Tests” might not follow the intuition of statistics. Likelihood Ratio criterion remains a generally reasonable first option for most problems, we still consider LRT to be a vital statistical tool.