DALARNA
UNIVERSITY

Data Collection and Data Quality

Lab – 1 Report

From

Gokul Kaisaravalli Bhojraj
Id: 80789
Program: Business Intelligence
e-mail: h18gokka@du.se

1. The data set "OhioSchool.csv" provides data on 1,965 Ohio Elementary School buildings for 2001-02 year.

a) Select a random 10 sample of 100 schools from this data set by using SRS (without replacement). Calculate the mean number of teacher per school. For each of the 10 samples. Do the sample estimates look good? How much do they vary?

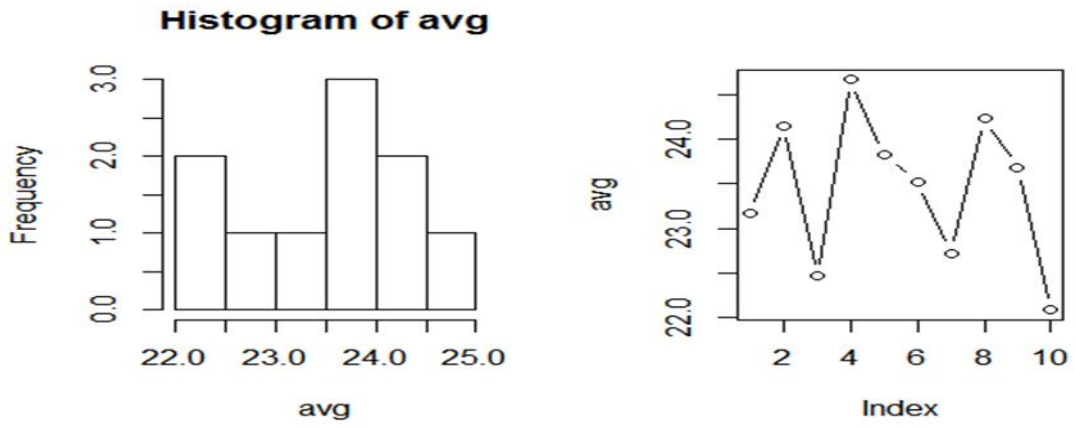
Solution:

Step 1: Select a sample of size 100 from Ohio School data without replacement randomly.
Step 2: Perform the above task for 10 times.
Step 3: Calculate the mean number of teachers each time.

Result:

- Population Mean: **24.07487**
- Overall mean number of teacher per school (10 times): **23.45**
- Mean number of teachers per school under each sampling:

23.17 24.15 22.46 24.67 23.82 23.52 22.72 24.23 23.68 22.08



- Sample average Estimate:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.08	22.83	23.60	23.45	24.07	24.67

As we can see the sample average estimate looks good. It ranges from 22.08 to 24.67 with mean b eing 23.45.

- Sample Estimate:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
6.00   16.00   23.00   24.33   29.00   65.00
```

As we can see the sample estimate. It ranges from 6 to 65 with mean being 24.3 teachers.

- Standard Deviation: **0.83332**

  This result indicates a standard deviation of **0.83332** meaning that the averages can spread far from the mean.

b) Select a sample of size 100 (or as closed to it as possible) by using the method of systematic random sampling. Describe your selection procedure. Also calculate the mean number teachers per school.

Solution:

Step 1: Since the data size is 1967 in order select 100 sample size, we need to have blocks of 20 (20*100=2000). So, we have to add 33 more.
Step 2: Select the row by taking first 33 or randomly take 1 column and it 33 times at the end.
Step 3: Once we have 2000 rows, select a number **n** randomly from 1 to 20.
Step 4: Pick the row after every **nth** rows.
Step 5: Calculate the average once the rows are chosen of total 100.

Result:

- Population Mean: **24.07487**
- Average number of teacher per school: **23.88**

c) Consider school district as clusters and calculate average number of teacher per school using a cluster sample (use 30 clusters).

Solution:

Step 1: list all the unique districts.
Step 2: Select 30 among those randomly.
Step 3: Select all the schools (rows) under those 30 districts (clusters).
Step 4: Calculate the average teacher per school from the sample.

Result:

- Population Mean: **24.07487**
- Average number of teacher per school: **23.65263**

d) Draw a stratified random sample of size 100 using poverty level as the strata (consider 4 quartiles). Also estimate the population mean of number of teacher per school.

Solution:

Step 1: Divide the whole data based on the poverty level considering the quantile values.

25%	50%	75%	100%
0.0611905	0.1192420	0.1935360	1.0000000

Step 2: Since the sample size is 100 and we have 4 strata, we need to draw sample from each strata proportional to the size of strata using the following formula

$$N_k = (\text{size of strata k}/\text{ size of whole population}) * \text{Overall sample size}$$

N: It is the no of samples to be drawn form that strata k.
Step 3: Take all the samples chosen from each strata.
Step 4: Calculate the average teacher per school from each strata sample.

Result:

- Population Mean: **24.07487**
- Average number of teacher per school: **24.77**

2. A survey is to be performed to determine a certain parameter in a community. A previous study showed population Sd being 46. If a sample error of up to 4 can be accepted, how many subjects should be included in this survey under 95% confidence.

Solution:

Formula:

$$n = Z_{\alpha}^{2}\frac{\sigma^{2}}{(\bar{x}-\mu)^{2}}$$

Where:

- n: Number of samples required.
- Z: It's the Confidence level, here it's 95% so value of Z is 1.96.
- σ: It is the Standard Deviation, here it is 46.
- (X-μ): Is the error accepted, here it is 4.

Result: We need to have at least **508** subjects for the survey to have 95% confidence.

3. Read Fan, et al. (1962), Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers, Journal of the American Statistical Association, 57(298): 387-402. Implement their algorithms 1-3, to draw a random sample of size 30 (schools) from Ohio School data.

Solution:

Note: As stated in the article "Since simple random sampling can be viewed as a special case of stratified random sampling where there is only one stratum, a separate presentation of simple random sampling plans will not be given".

Assumption: Based on the above statement the whole Ohio School data is considered as a single Strata and the based on this the following algorithms are implemented.

- **Algorithm 1: Stratified Random Sampling Plan:**

Select a sample of size 30 from Ohio School data (size known) without replacement randomly.

R code:

```
> All_rows<-seq(1,length(data$dist_irn),1)
> picked_rows<-sample(All_rows,size=30,replace=FALSE)
> picked_data<-data.frame(data[picked_rows,])
> nrow(picked_data)
[1] 30
```

Result:

As we can see the total rows picked are 30 as required.

If we try to use the samples to find the average teachers in each school, it is as follows

- Population Mean: **24.07487**
- Average number of teacher per school: **24.56667**

➢ **Algorithm 2: Segmentized Sampling Plan:**

Step 1: Divide the whole data based on the poverty level considering the quantile values.

```
      25%       50%       75%      100%
0.0611905 0.1192420 0.1935360 1.0000000
```

Step 2: Since the sample size is 30 and we have 4 strata, we need to draw sample from each strata proportional to the size of strata using the following formula

N_k = (size of strata k/ size of whole population) * Overall sample size

N: It is the no of samples to be drawn form that strata k.

Step 3: Take all the samples chosen from each strata.

Step 4: Calculate the average teacher per school from each strata sample.

R code:

```
> quantile(data$poverty,c(0.25,0.50,0.75,1))

      25%       50%       75%      100%
0.0611905 0.1192420 0.1935360 1.0000000

> q_25<-0.0611905
> q_50<-0.1192420
> q_75<-0.1935360
> q_100<-1.0000000
>
> poverty_below_25 <- data[data$poverty<=q_25,]
> nrow(poverty_below_25)
[1] 492
>
> poverty_below_50_a <- subset(data,data$poverty>q_25)
> nrow(poverty_below_50_a)
[1] 1475
> poverty_below_50<-subset(poverty_below_50_a,poverty_below_50_a$poverty<=q_50)
> nrow(poverty_below_50)
[1] 494
>
> poverty_below_75_a<- subset(data,data$poverty>q_50)
> nrow(poverty_below_75_a)
[1] 981
> poverty_below_75<-subset(poverty_below_75_a,poverty_below_75_a$poverty<=q_75)
> nrow(poverty_below_75)
[1] 490
```

```
> poverty_below_100<-subset(data,data$poverty>q_75)
> nrow(poverty_below_100)
[1] 491
>
> nrow(poverty_below_25)+nrow(poverty_below_50)+nrow(poverty_below_75)+nrow(poverty_below_100)
[1] 1967
>
>
> drawing_size<-(nrow(poverty_below_25)/nrow(data))*30
> All_rows<-seq(1,length(poverty_below_25$dist_irn),1)
> picked_row_strata_1<-sample(All_rows,size=round(drawing_size),replace=FALSE)
> picked_data_strata_1<-data.frame(poverty_below_25[picked_row_strata_1,])
>
> drawing_size<-(nrow(poverty_below_50)/nrow(data))*30
> All_rows<-seq(1,length(poverty_below_50$dist_irn),1)
> picked_row_strata_2<-sample(All_rows,size=round(drawing_size),replace=FALSE)
> picked_data_strata_2<-data.frame(poverty_below_50[picked_row_strata_2,])
>
> drawing_size<-(nrow(poverty_below_75)/nrow(data))*30
> All_rows<-seq(1,length(poverty_below_75$dist_irn),1)
> picked_row_strata_3<-sample(All_rows,size=round(drawing_size),replace=FALSE)
> picked_data_strata_3<-data.frame(poverty_below_75[picked_row_strata_3,])
>
> drawing_size<-(nrow(poverty_below_100)/nrow(data))*30
> All_rows<-seq(1,length(poverty_below_25$dist_irn),1)
> picked_row_strata_4<-sample(All_rows,size=round(drawing_size),replace=FALSE)
> picked_data_strata_4<-data.frame(poverty_below_100[picked_row_strata_4,])
>
> all_sample<-rbind(picked_data_strata_1,picked_data_strata_2,
+                   picked_data_strata_3,picked_data_strata_4)
>
>
> nrow(all_sample)
[1] 30
```

Result:

As we can see the total rows picked are 30 as required.

If we try to use the samples to find the average teachers in each school, it is as follows

- Population Mean: **24.07487**
- Average number of teacher per school: **26.6**

- **Algorithm 3: Systematic Sampling Plan:**

Step 1: Since the data size is 1967 in order select 30 sample size, we need to have blocks of 66 (66*30=1980).

Step 2: select a number **n** randomly from 1 to 66.

Step 3: Pick the row after every **nth** rows.

R code:

```
> Row_div<-length(data$ZIP)/30
> ceiling(Row_div)
[1] 66
>
> frame_1<-cbind(row=c(1:nrow(data)),seq=rep(1:66,30))
>
> select<- sample(1:66,1)
> sample_2<-frame_1[frame_1[,2]==select,]
> sys1<-data[sample_2[,1],c("build_irn","teachers")]
> nrow(sys1)
[1] 30
```

Result:

As we can see the total rows picked are 30 as required.

If we try to use the samples to find the average teachers in each school, it is as follows

- Population Mean: **24.07487**
- Average number of teacher per school: **24.3**

****