MA5750: APPLIED STATISTICS
ASSIGNMENT 1

Monte Carlo Exercises

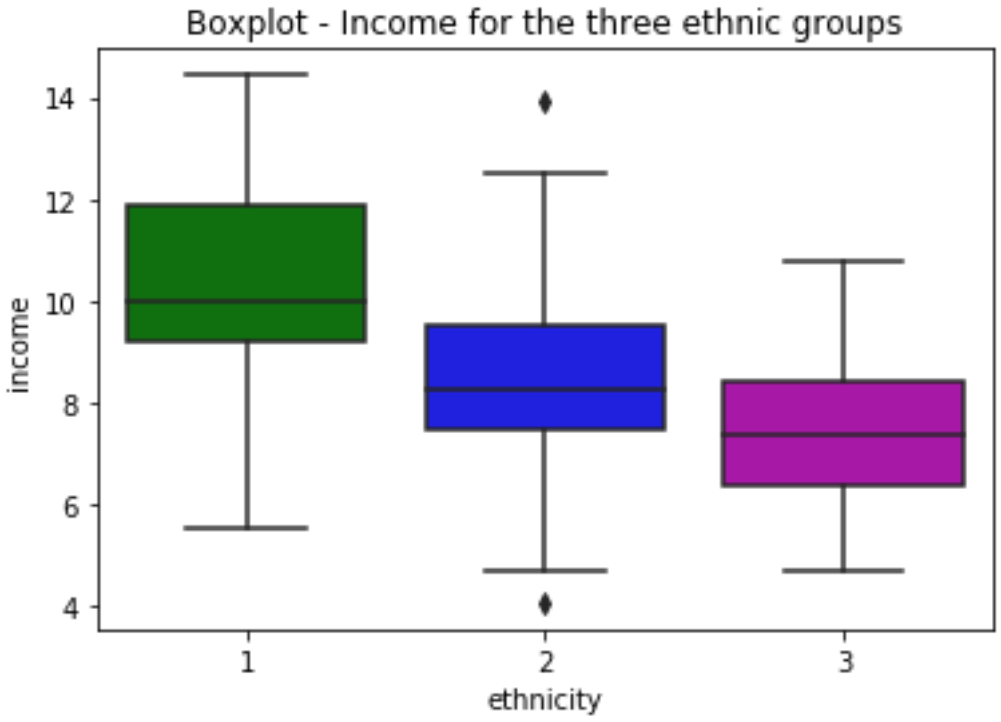
2.1.

(a)

```
1. import pandas as pd
2. import numpy as np
3. from collections import Counter
4. import seaborn as sns
```

```
5. df = pd.read_csv('H:/J - Applied Statistics/Assignments/1/sscsample.csv')

6. my_pal = {1: "g", 2: "b", 3:"m"}
7. sns.boxplot( x=df["ethnicity"], y=df["income"], palette=my_pal).set_title('Boxplot - Incomes for the three ethnic groups')
```




```
1. for i in range(1,4):
2.     print('Mean income for ethnicity', i, 'is', df[df['ethnicity']==i]['income'].mean())
```

```
Mean income for ethnicity 1 is 10.3104375
Mean income for ethnicity 2 is 8.410405
Mean income for ethnicity 3 is 7.529680000000002
```

The mean income for ethnicity 1 is the highest followed by ethnicity 2 and the by ethnicity 3. There is an inherent difference in the average value of incomes in different ethnicity.

(b)

i. Simple Random Sampling does not always have the strata represented in the correct proportions.

Because, we are sampling at random from the entire population and nowhere do we inherently exercise control in selection based on ethnicity at every instant of sampling.

ii. Yes, on an average Simple Random Sampling give the strata in their correct proportions.

As the sampling is Uniform Random and with the proportion in ethnicity in the population, we can expect on an average that the proportion be present in the sample as well. Say, if a particular ethnicity is greater in number then there is a greater chance that a datapoint belonging to that ethnicity lands up in the

sample and vice-versa. Over many samples Simple Random Sampling give the strata in their correct proportions on average.

iii. The Population mean is 8.994273

```
# Simple Random Sampling
np.random.seed(0)
mean_income_RandomSampling = []
for i in range(200):
    index = np.random.randint(low=1, high=101, size=20)
    df_temp = df.loc[index]
    mean_income_RandomSampling.append(df_temp['income'].mean())

print("Mean of the sampling distribution of the sample mean for simple random sampling:"
      , sum(mean_income_RandomSampling)/len(mean_income_RandomSampling))
```

Mean of the sampling distribution of the sample mean for simple random sampling: 8.994810428242173

Yes, it appears to be close to the population mean and the difference may be due to chance alone as we are taking only 200 samples. The value tends to the population mean on taking infinite samples.

(c)

i. Stratified Random Sampling always have the strata represented in the correct proportions.

Because, we are sampling at random from each subpopulation in their corresponding proportions.

ii. Yes, on an average Stratified Random Sampling give the strata in their correct proportions.

When Stratified Random Sampling always have the strata represented in the correct proportions it should be the same on an average as well.

iii. The Population mean is 8.994273

```
# Stratified Random Sampling
np.random.seed(0)
mean_income_StratifiedRandomSampling = []
index_dict = [{'low':1,'high':41},
              {'low':41,'high':81},
              {'low':81,'high':101}
             ]
index_prop = [40*20/100, 40*20/100, 20*20/100]
for i in range(200):
    index = []
    for j in range(3):
        for x in list(np.random.randint(low=index_dict[j]['low'], high=index_dict[j]['high'], size=int(index_prop[j]))):
            index.append(x)
    df_temp = df.loc[index]
    mean_income_StratifiedRandomSampling.append(df_temp['income'].mean())

print("Mean of the sampling distribution of the sample mean for Stratified random sampling:"
      , sum(mean_income_StratifiedRandomSampling)/len(mean_income_StratifiedRandomSampling))
```

Mean of the sampling distribution of the sample mean for Stratified random sampling: 8.975331453947366

Yes, it appears to be close to the population mean and the difference may be due to chance alone as we are taking only 200 samples. The value tends to the population mean on taking infinite samples.

(d)

i. Cluster Random Sampling does not always have the strata represented in the correct proportions.

Because, we are sampling at random a set of clusters and each cluster may have different proportions of ethnicity.

ii. Yes, on an average Cluster Random Sampling give the strata in their correct proportions.

iii. The Population mean is 8.994273

```
# Cluster Random Sampling
np.random.seed(0)
mean_income_ClusterRandomSampling = []

for i in range(200):
    incomes = []
    for j in np.random.randint(low=1, high=21, size=6):
        for income in df[df['neighborhood']==j]['income'].values:
            incomes.append(income)
    mean_income_ClusterRandomSampling.append(sum(incomes[:20])/20)

print("Mean of the sampling distribution of the sample mean for Cluster random sampling:"
      , sum(mean_income_ClusterRandomSampling)/len(mean_income_ClusterRandomSampling))
```

Mean of the sampling distribution of the sample mean for Cluster random sampling: 9.025136124999996

(e)

```
# Spreads of  the  sampling distributions
mean_incomes = {'ClusterRandomSampling':mean_income_ClusterRandomSampling,
                'StratifiedRandomSampling':mean_income_StratifiedRandomSampling,
                'RandomSampling':mean_income_RandomSampling}
for method in mean_incomes:
    print(method,':')
    print('Standard Deviation =',statistics.stdev(mean_incomes[method]))
    print('Inter Quartile Range =',iqr(mean_incomes[method]))
```

RandomSampling :
Standard Deviation = 0.5082238642647554
Inter Quartile Range = 0.73581875

StratifiedRandomSampling :
Standard Deviation = 0.46517088794762035
Inter Quartile Range = 0.628253749999999

ClusterRandomSampling :
Standard Deviation = 0.7694465714668365
Inter Quartile Range = 1.0226875

From the Standard Deviation and Inter Quartile Range Stratified Random Sampling seems to be more effective in giving sample means more concentrated about the true mean.

(f) This is because in Stratified Random Sampling we sample proportionally based on ethnicity each and every time and hence the variance (or standard deviation) and IQR is expected to be small.

2.2.

(a)

i. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design with a little bit of standard deviation.

ii. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a randomized block design with smaller standard deviation than completely randomized design.

iii. No the distribution varies in two designs. In Randomized block design the Treatment groups always contain the same distribution of other variable whereas in Completely Randomized design it is not always the case and only on an average are they same.

iv. Randomized block design is controlling for the other variable more effectively due to its inherent method of forming blocks based on other variable prior to randomly assigning to each of treatment groups as opposed to Completely Randomized design where the other variable is not considered and the units are assigned randomly entirely.

v. No, on average all groups do not have the same underlying mean value for the response variable when we use a completely randomized design. This is because we do not have complete control over the lurking variable which in this case has high correlation with Response variable.

vi. Yes, on average all groups have the same underlying mean value for the response variable when we use a randomized block design. Even if the other variable has high correlation with response variable it is being taken care that the distribution of other variable is same across groups.

vii. The distribution of the Response variable might be different over different designs for a given treatment group because of the varying distribution of Other variable even if the distribution of Treatment effect (T_j) is same.

viii. Randomized Block design will give us a better chance for detecting a small difference in treatment effects. This is because the distribution of other variable is controlled over the different groups to be similar. In Completely Randomized design even small changes in the distribution due to chance will affect the Response variable, due to high correlation, on top of the small difference in treatment effects.

ix. Yes, blocking on the other variable is effective when the response variable is strongly related to the other variable. By controlling the distribution of other variable, we can ensure that the effect due to it on Response variable is consistent.

(b)

i. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design with a little bit of standard deviation.

ii. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a randomized block design with smaller standard deviation than completely randomized design.

iii. No the distribution varies in two designs. In Randomized block design the Treatment groups always contain the same distribution of other variable whereas in Completely Randomized design it is not always the case and only on an average are they same.

iv. Randomized block design is controlling for the other variable more effectively due to its inherent method of forming blocks based on other variable prior to randomly assigning to each of treatment groups as opposed to Completely Randomized design where the other variable is not considered and the units are assigned randomly entirely.

v. No, on average all groups do not have the same underlying mean value for the response variable when we use a completely randomized design. But it will be closer to the mean value now since the lurking variable has low correlation with Response variable. Randomized Block Design will now provide only slighter edge over the other design.

vi. Yes, on average all groups have the same underlying mean value for the response variable when we use a randomized block design. Even if the other variable has high correlation with response variable it is being taken care that the distribution of other variable is same across groups.

vii. The distribution of the Response variable might be different over different designs for a given treatment group because of the varying distribution of Other variable even if the distribution of Treatment effect (T_j) is same. But in this case, we can expect it to be almost appear to be same as the correlation between Other variable and Response Variable is small. Randomized Block Design will now provide only slighter edge over the other design.

viii. Randomized Block design will give us a better chance for detecting a small difference in treatment effects. This is because the distribution of other variable is controlled over the different groups to be similar. In Completely Randomized design even small changes in the distribution due to chance will affect the Response variable but in a little way, due to low correlation, on top of the small difference in treatment effects.

ix. Yes, blocking on the other variable is effective when the response variable is strongly related to the other variable. By controlling the distribution of other variable, we can ensure that the effect due to it on Response variable is consistent. But is less effective when the response variable is weakly related to the other variable.

(c)

i. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design with a little bit of standard deviation.

ii. Yes, on average we can say all groups have the same underlying mean value for the other (lurking) variable when we use a randomized block design with smaller standard deviation than completely randomized design.

iii. No the distribution varies in two designs. In Randomized block design the Treatment groups always contain the same distribution of other variable whereas in Completely Randomized design it is not always the case and only on an average are they same.

iv. Randomized block design is controlling for the other variable more effectively due to its inherent method of forming blocks based on other variable prior to randomly assigning to each of treatment groups as opposed to Completely Randomized design where the other variable is not considered and the units are assigned randomly entirely.

In this case where the correlation is zero none of the methods provide any relative advantage of controlling the other variable more effectively. Hence both are equal and in fact provide no control

v. Yes, on average all groups have the same underlying mean value for the response variable when we use a completely randomized design. As there is no correlation between other variable and response variable it does not matter how we choose the units as none of them have relative effect on response variable.

vi. Yes, on average all groups have the same underlying mean value for the response variable when we use a randomized block design.

vii. The distribution of the Response variable is same over different designs for a given treatment group because the varying distribution of Other variable has no effect on response variable.

viii. In this case no design will give better chance of detecting small difference in treatment effects as Randomized block design does not provide the edge over Completely randomized design due to zero correlation between other variable and response variable.

ix. Naturally blocking on the other variable is not effective when the response variable is independent from the other variable.

x. We will not lose any effectiveness to the end result of the Experimentation but it is important to note that Randomized Block design is no better than Completely randomized design and we could save the energy on creating blocks based on the other variable which is unnecessary. Relying on Completely randomized design is more efficient in this case.