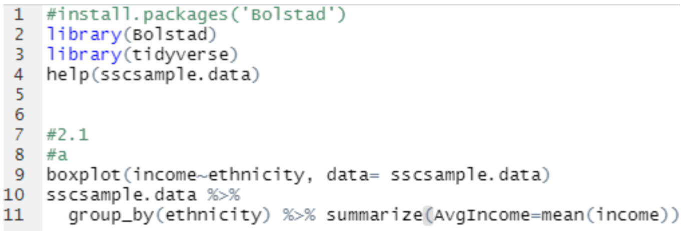
2.1.

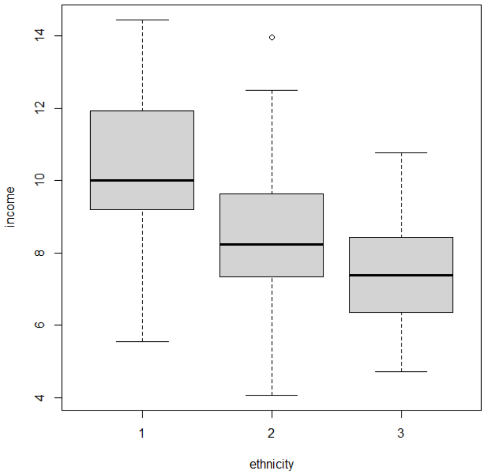
a. **Display the incomes for the three ethnic groups (strata) using boxplots on the same scale. Compute the mean income for the ethnic groups. Do you see any difference between the income distributions?**

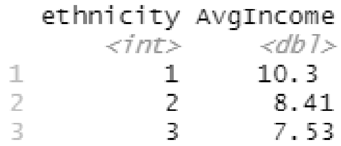
The mean income for group 1 is $103,000.
The mean income for group 2 is $84,100.
The mean income for group 3 is $75,300.

Yes, the distributions are different. Ethnicity 1 has a wide spread of incomes from just under $60,000 to over $150,000. Ethnicity 2 is also wide with a spread from just over $40,000 to nearly $130,000. Ethnicity 3 has a much smaller range of incomes with a spread from roughly $50,000 to around $110,000.



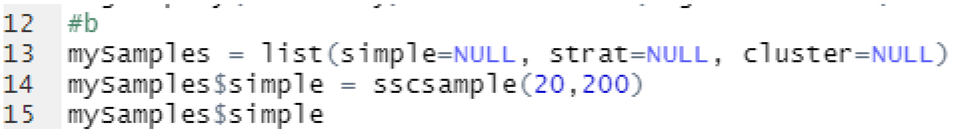
```
#install.packages('Bolstad')
library(Bolstad)
library(tidyverse)
help(sscsample.data)


#2.1
#a
boxplot(income~ethnicity, data= sscsample.data)
sscsample.data %>%
  group_by(ethnicity) %>% summarize(AvgIncome=mean(income))
```



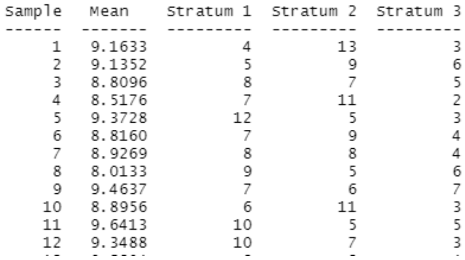


```
  ethnicity AvgIncome
      <int>     <dbl>
1         1     10.3 
2         2      8.41
3         3      7.53
```

b. **Draw 200 random samples of size 20 from the population using the sscsample function.**



```
#b
mySamples = list(simple=NULL, strat=NULL, cluster=NULL)
mySamples$simple = sscsample(20,200)
mySamples$simple
```

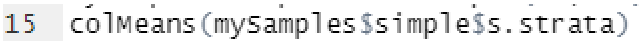


```
Sample    Mean    Stratum 1  Stratum 2  Stratum 3
------   -------  ---------  ---------  ---------
     1   9.1633          4         13          3
     2   9.1352          5          9          6
     3   8.8096          8          7          5
     4   8.5176          7         11          2
     5   9.3728         12          5          3
     6   8.8160          7          9          4
     7   8.9269          8          8          4
     8   8.0133          9          5          6
     9   9.4637          7          6          7
    10   8.8956          6         11          3
    11   9.6413         10          5          5
    12   9.3488         10          7          3
```

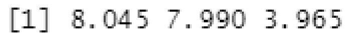
i. **Does the simple random sampling always have the strata represented in the correct proportions?**

No, as you can see in the output above, the proportions are different in the samples. They do not hold up to the 40, 40, 20 split they are supposed to have.

ii. **On the average, does simple random sampling give the strata their correct proportions?**



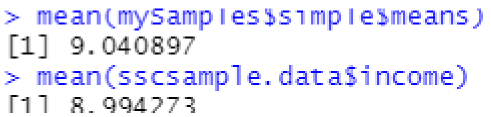
```
colMeans(mySamples$simple$s.strata)
```



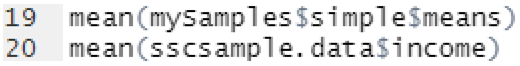
```
[1] 8.045 7.990 3.965
```

Technically, no. However, based on the average of each, they are close enough to 8, 8, 4 (40%, 40%, 20%) that I would say it does give them their correct proportions on average.

iii. **Does the mean of the sampling distribution of the sample mean for the simple random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone?**



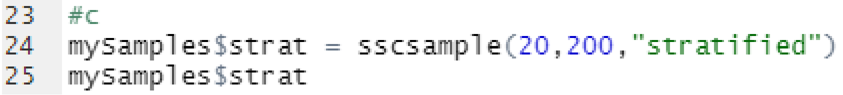
```
> mean(mySamples$simple$means)
[1] 9.040897
> mean(sscsample.data$income)
[1] 8.994273
```



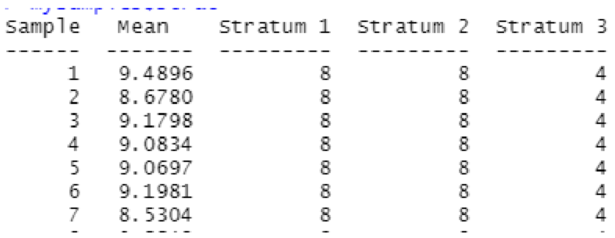
```
mean(mySamples$simple$means)
mean(sscsample.data$income)
```

I would say it is close enough to be caused by chance. The difference between means is less than 0.05. If we had more samples, then the number probably gets closer to the population mean.

c. **Draw 200 stratified random samples using the function and store the output in mySample$strat**



```
#c
mySamples$strat = sscsample(20,200,"stratified")
mySamples$strat
```

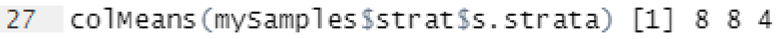


```
Sample   Mean    Stratum 1  Stratum 2  Stratum 3
------  -------  ---------  ---------  ---------
     1   9.4896          8          8          4
     2   8.6780          8          8          4
     3   9.1798          8          8          4
     4   9.0834          8          8          4
     5   9.0697          8          8          4
     6   9.1981          8          8          4
     7   8.5304          8          8          4
```

i. **Does the stratified random sampling always have the strata represented in the correct proportions?**

Yes, as you can see in the output above, the proportions are the same in each sample. They do hold up to the 40, 40, 20 split they are supposed to have.

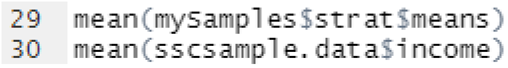
ii. **On the average, does stratified random sampling give the strata their correct proportions?**



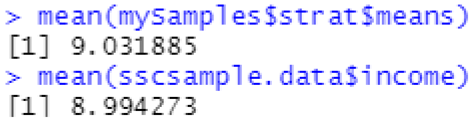
```
colMeans(mySamples$strat$s.strata) [1] 8 8 4
```

Yes, the proportions are exactly right on average for stratified random sampling. They are exactly 8, 8, 4 (40%, 40%, 20%).

iii. **Does the mean of the sampling distribution of the sample mean for the stratified random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone?**



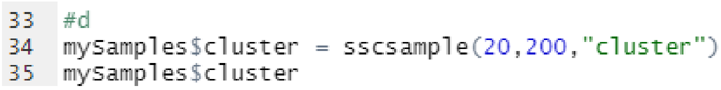
```
mean(mySamples$strat$means)
mean(sscsample.data$income)
```



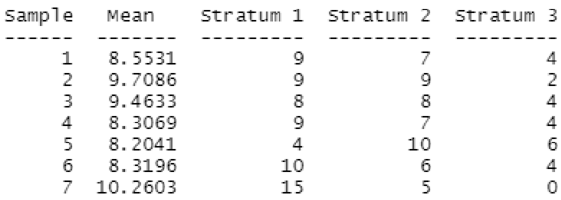
```
> mean(mySamples$strat$means)
[1] 9.031885
> mean(sscsample.data$income)
[1] 8.994273
```

I would say it is close enough to be caused by chance. The difference between means is less than 0.04. If we had more samples, then the number probably gets even closer to the population mean.

d. **Draw 200 cluster random samples using the function and store the output in mySamples$cluster**



```
#d
mySamples$cluster = sscsample(20,200,"cluster")
mySamples$cluster
```

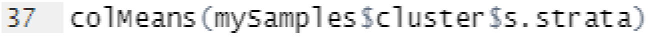


```
Sample   Mean    Stratum 1  Stratum 2  Stratum 3
------  -------  ---------  ---------  ---------
     1   8.5531          9          7          4
     2   9.7086          9          9          2
     3   9.4633          8          8          4
     4   8.3069          9          7          4
     5   8.2041          4         10          6
     6   8.3196         10          6          4
     7  10.2603         15          5          0
```

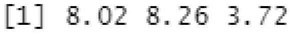
i. **Does the cluster random sampling always have the strata represented in the correct proportions?**

No, as you can see in the output above, the proportions are different in the samples. They do not hold up to the 40, 40, 20 split they are supposed to have.

ii. **On the average, does cluster random sampling give the strata their correct proportions?**



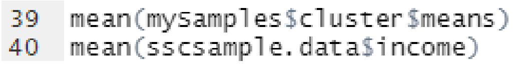
```
colMeans(mySamples$cluster$s.strata)
```



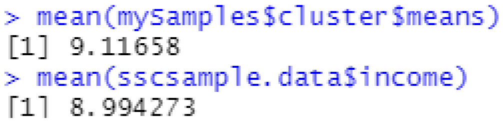
```
[1] 8.02 8.26 3.72
```

Technically, no. However, based on the average of each strata, they are close enough to 8, 8, 4 (40%, 40%, 20%) that I would say it does give them their correct proportions on average.

iii. **Does the mean of the sampling distribution of the sample mean for the cluster random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone?**



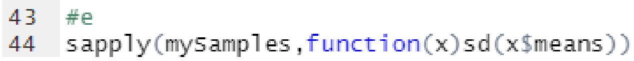
```
mean(mySamples$cluster$means)
mean(sscsample.data$income)
```



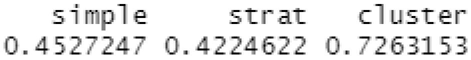
```
> mean(mySamples$cluster$means)
[1] 9.11658
> mean(sscsample.data$income)
[1] 8.994273
```

I would say it is close enough to be caused by chance. The difference between means is just over 0.12. Which would be about $1,200. If we had more samples, then the number probably gets closer to the population mean.

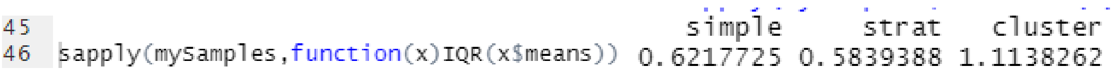
e. **Compare the spreads of the sampling distributions (stan dev and interquartile range), which method of random sampling seems more effective in giving simple means more concentrated about the true mean?**



```
#e
sapply(mySamples,function(x)sd(x$means))
```



```
   simple     strat   cluster 
0.4527247 0.4224622 0.7263153 
```



```
sapply(mySamples,function(x)IQR(x$means))
```

```
    simple     strat   cluster 
0.6217725 0.5839388 1.1138262 
```

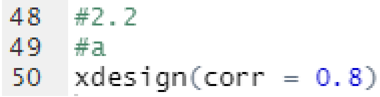
The stratified random sampling method seems more effective in giving simple means concentrated about the true mean.

f. Give reasons for this.

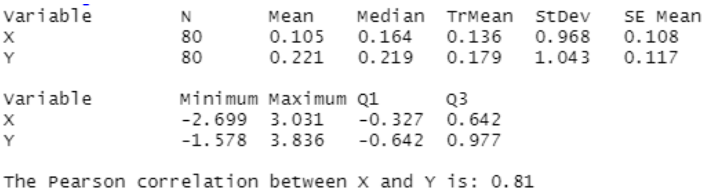
I think this because it has the lowest standard deviation of means and the lowest interquartile range. Based on these figures, it is also safe to say that cluster is the worst of the three methods.

2.2

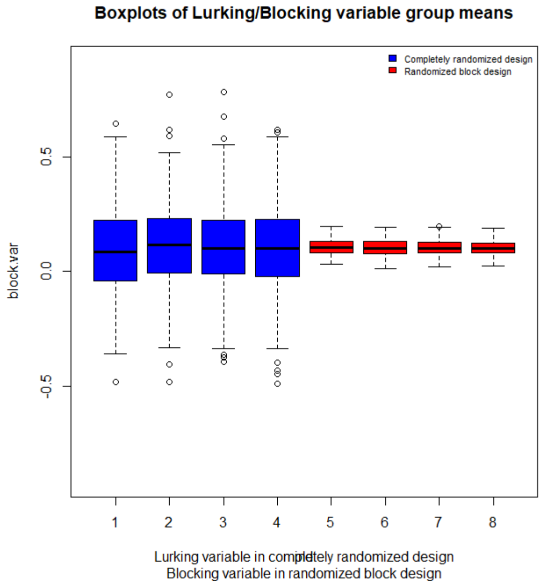
a. First, we will do a small-scale Monte Carlo study of 500 random assignments using each of the two designs when the response variable is strongly related to the other variable. We let the correlation between them be ρ = 0.8.

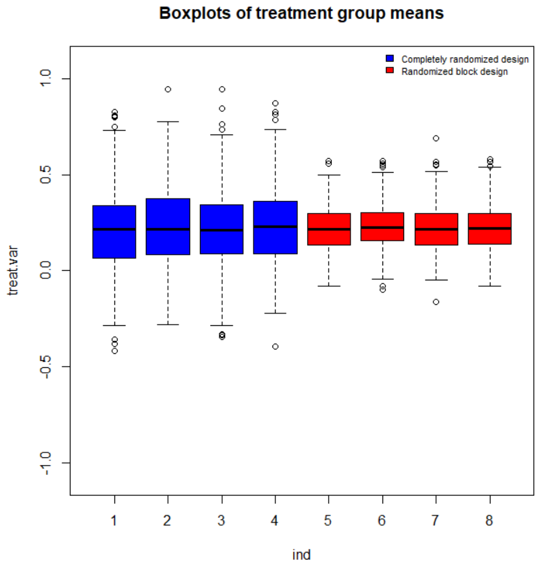


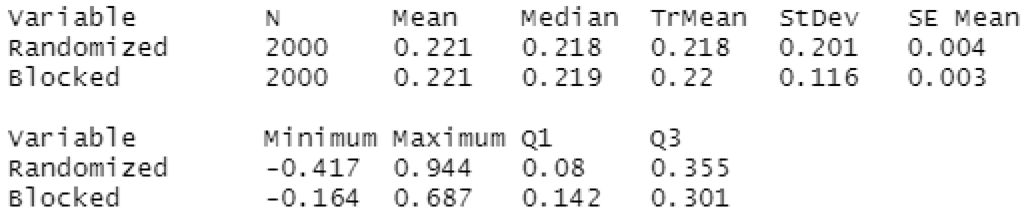
```
#2.2
#a
xdesign(corr = 0.8)
```



```
Variable         N        Mean     Median   TrMean   StDev    SE Mean
X                80       0.105    0.164    0.136    0.968    0.108
Y                80       0.221    0.219    0.179    1.043    0.117

Variable         Minimum  Maximum  Q1       Q3
X                -2.699   3.031    -0.327   0.642
Y                -1.578   3.836    -0.642   0.977

The Pearson correlation between X and Y is: 0.81
```







```
Variable         N        Mean     Median   TrMean   StDev    SE Mean
Randomized       2000     0.221    0.218    0.218    0.201    0.004
Blocked          2000     0.221    0.219    0.22     0.116    0.003

Variable         Minimum  Maximum  Q1       Q3
Randomized       -0.417   0.944    0.08     0.355
Blocked          -0.164   0.687    0.142    0.301
```

i. Does it appear that on average, all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design?

I would say yes, based on the box plots.

ii. **Does it appear that on average, all groups have the same underlying mean value for the other (blocking) variable when we use a randomized design?**
I would say yes, based on the box plots.

iii. **Does the distribution of the other variable over the treatment groups appear to be the same for the two designs? Explain any difference**
No, the completely randomized design has a much wider range in comparison to the randomized block design.

iv. **Which design is controlling for the other variable more effectively? Explain.**
The block design because it had a tighter spread

v. **Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a completely randomized design.**
Yes, because the means were about the same

vi. **Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a randomized block design.**
Yes, because the means were about the same

vii. **Does the distribution of the response variable over the treatment groups appear to be the same for the two designs? Explain any difference.**
No, the completely randomized design has a much wider range in comparison to the randomized block design.

viii. **Which design will give us a better chance for detecting a small difference in treatment effect? Explain**
The block design because it had a tighter spread

ix. **Is blocking on the other variable effective when the response is strongly related to the other variable.**
I would say so. The range appears much smaller and therefore more accurate.