EDA A3 Q3: Judgment sampling

Consider a study population $\mathcal{P}_{Study}$ consisting of $N = 100$ blocks labelled $u = 1, 2, 3, \ldots, 100$.

The blocks are of of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5mm thickness), but have different shapes. Suppose also that $\mathcal{P}_{Target} = \mathcal{P}_{Study}$ and that the population attribute of interest $a(P_{Target}) = \frac{1}{N}\sum_{u \in P_{Target}} weight(u)$ that is the average weight of all $N = 100$ blocks in the population.

We want a sample $\mathcal{S} \subset \mathcal{P}_{Study}$ of $n = 10$ blocks selected from the 100, whose average weight is (nearly) the same as the average weight of all 100.

That is, we would like a sample with zero (or at least small in absolute value) **sample error** $a(\mathcal{S}) - a(\mathcal{P}_{Study})$.

The 'blocks' data can be loaded from the assignment data directory as follows:

```
load("/Users/rudranibhadra/Downloads/blocks.rda")
head(blocks, n = 3)
```

```
##   id weight perimeter group
## 1  1     55        32     B
## 2  2     35        27     B
## 3  3     35        25     A
```

```
#blocks
```

Having been presented with all 100 blocks and asked to **judge** which 10 blocks have an average weight nearest the average weight of all 100 blocks, each student would have come up with their own sampling plan based on their judgment. This type of sampling is called **judgment sampling**.

The id numbers of the students and the blocks they selected are recorded in another file, `judgmentSamples.csv`. These can be loaded from the assignment data directory as follows:

```
students <- read.csv("/Users/rudranibhadra/Downloads/judgmentSamples.csv")
head(students)
```

```
##   studentID first second third fourth fifth sixth seventh eighth ninth
## 1      5086    12     18    17     11    15    20      14     13    16
## 2      3848    34     35    70     56    32    14       5     88    81
## 3      6656    14     34    41     29    32    55      74     40    16
## 4      7548    38     25    21     36    39    95      76     40    41
## 5      4114    66     37    31     44    94     2       8     51    23
## 6      9470    97     16    44     92    73    25      38     22    91
##   tenth
## 1    18
## 2    73
## 3    70
## 4    64
## 5    91
## 6    43
```

The variates of 'student' identify the student and the id numbers of the blocks they selected, in the order they recorded them.

a. *(4 marks)* Draw a histogram of all of the block weights selected by the students. If any block was selected by more than one student, include its weight as often as it was selected. That is, there will be a total of 'r nrow(students) * (ncol(students) - 1)' weights used to construct the histogram.

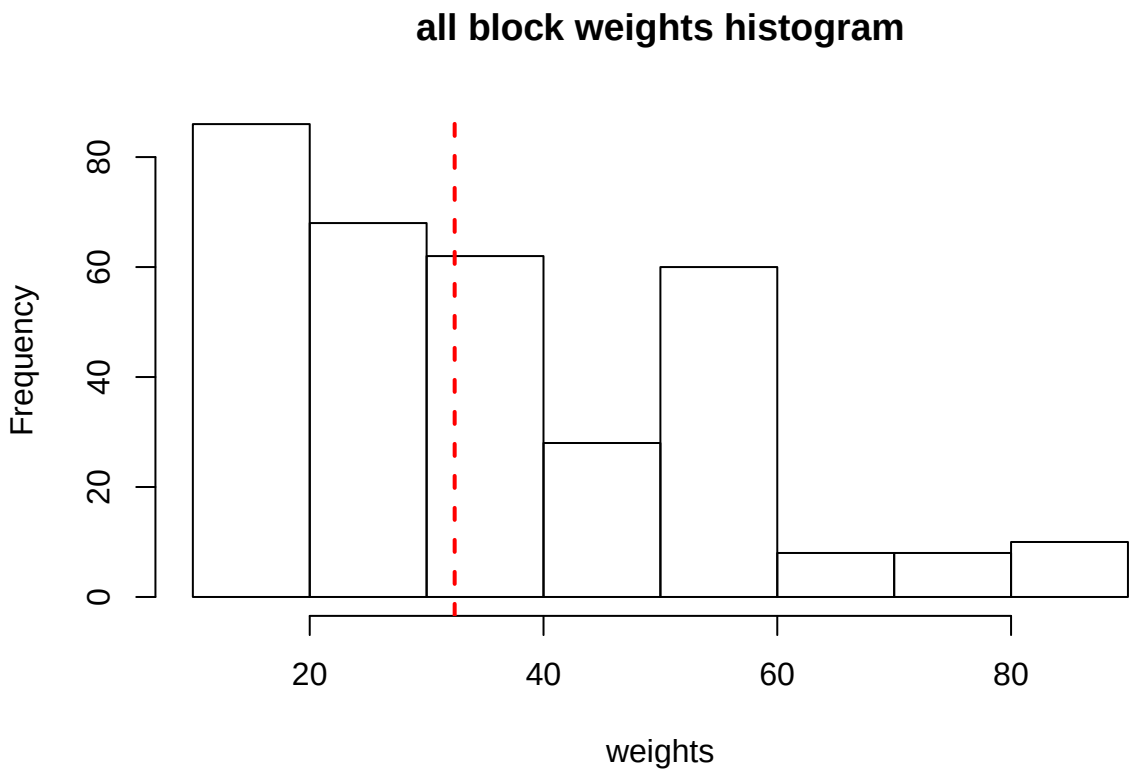
- Make sure the histogram is suitably labelled

- Add a vertical dashed red line at the at the average of all 100 weights in the entire population of 100 blocks (i.e. not just those selected by students).

Show your code.

```
s<-students
s[]<-lapply(students,function(x) blocks$weight[match(x,blocks$id)])
df<-cbind(students[,1],s[,-1])
#head(df)

hist(c(df$first,df$second,df$third,df$fourth, df$fifth,df$sixth,df$seventh,df$eighth,df$ninth, df$tenth
#x<-mean(blocks$weight)
abline(v=mean(blocks$weight),col="red", lwd=2, lty=2)
```



b. *(5 marks)* For each student, calculate the sample average weight of the blocks they selected. Create a data frame called 'judgmentErrors' of the student ids and their sample errors. Print out the ids and sample errors for both the top five and the bottom five students in increasing order of their *absolute* sample error.

Show your code.

```
x<-mean(blocks$weight)
judgementErrors<-data.frame(studentID=df[,1],SampleError=(rowMeans(df[,-1])-x))

#top 5
h<-(head(judgementErrors[with(judgementErrors,order(abs(judgementErrors$SampleError),
h[with(h,order(abs(h$SampleError),decreasing=FALSE)),]
```

```
##    studentID SampleError
```

```
## 1       5086        11.6
## 3       6656        11.6
## 22      7231        12.1
## 5       4114        12.6
## 27      7582        12.6
```

```
#bottom 5
head(judgementErrors[with(judgementErrors,order(abs(judgementErrors$SampleError),
decreasing=FALSE)),],n=5)
```

```
##    studentID SampleError
## 14      7656         2.6
## 17      7626         2.6
## 31      8395         2.6
## 12       842         3.1
## 26      7954        -3.4
```

c. *(3 marks)* Estimate the sampling bias and the sampling standard deviation for judgment sampling on this data. Show your code.

```
#sampling bias
n<-nrow(judgementErrors)
sum(judgementErrors$SampleError)/n
```

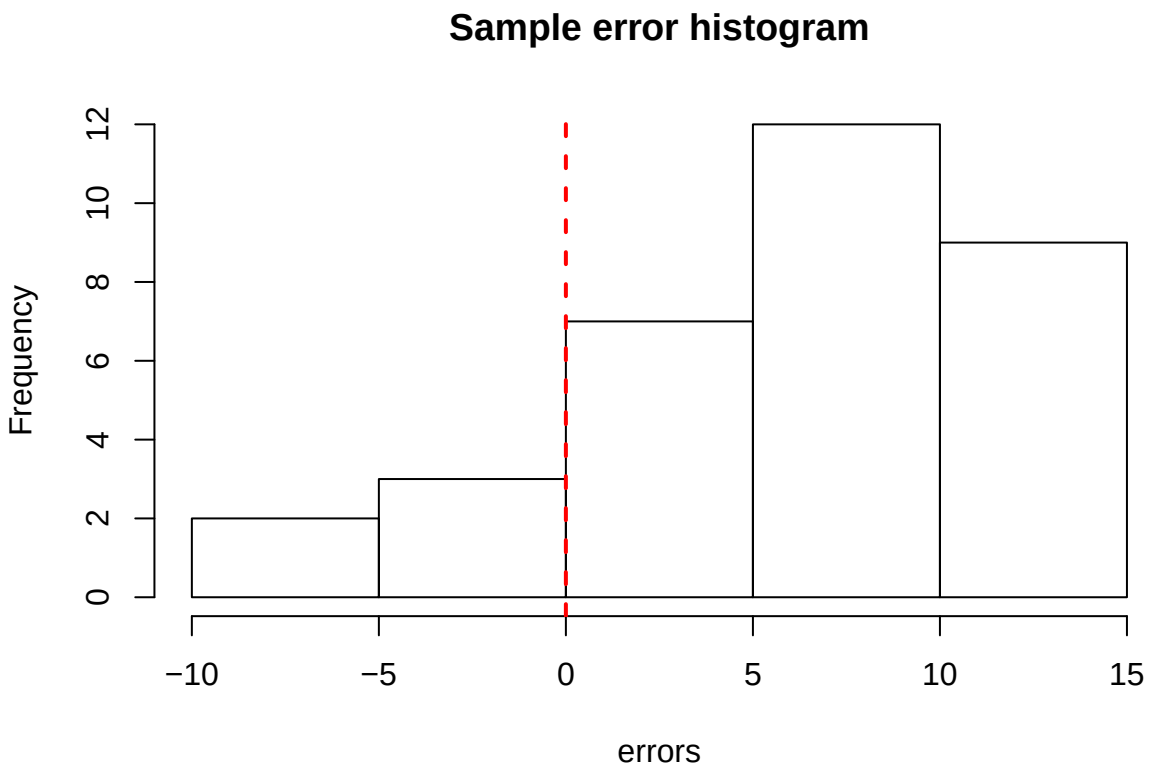
```
## [1] 5.418182
```

```
#sampling standard deviation
s<-rowMeans(df[,-1])
sd(s)
```

```
## [1] 5.508258
```

d. *(3 marks)* Provide a (suitably labelled) histogram of the sample errors. Add a vertical red dashed line at 0.

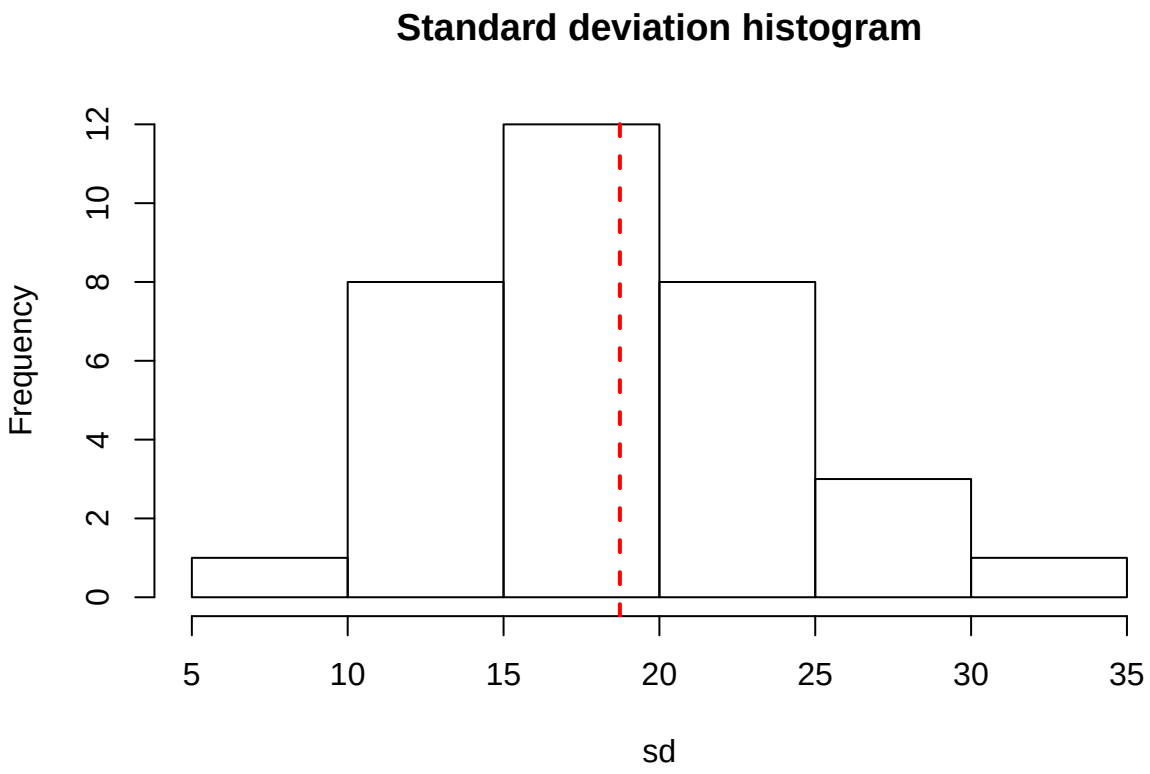
```
hist(judgementErrors$SampleError,main="Sample error histogram",xlab="errors")
abline(v=0, col="red",lwd=2,lty=2)
```



e. *(3 marks)* Calculate the sample standard deviation of the weights selected for each of the judgment samples. Draw a histogram of these standard deviations (suitably labelled). Draw a vertical dashed red line at the average of these standard deviations.

Show your code.

```
d1<-transform(df,SD=apply(df[,-1],1,sd,na.rm=TRUE))
hist(d1$SD,main="Standard deviation histogram",xlab="sd")
abline(v=mean(d1$SD),col="red", lwd=2, lty=2)
```



f. *(6 marks)* Identify which student had the smallest sample standard deviation **and** which student had the largest sample standard deviation. Report their standard deviations. Draw histograms (suitably labelled **and** having the same 'xlim = extendrange(blocks$weight)') of the weights of the blocks selected by each of these students. Add a vertical dashed red line to each histogram at the average of all 100 block weights in the population. What do you conclude about the sampling plan of each of these students?

Show your code.

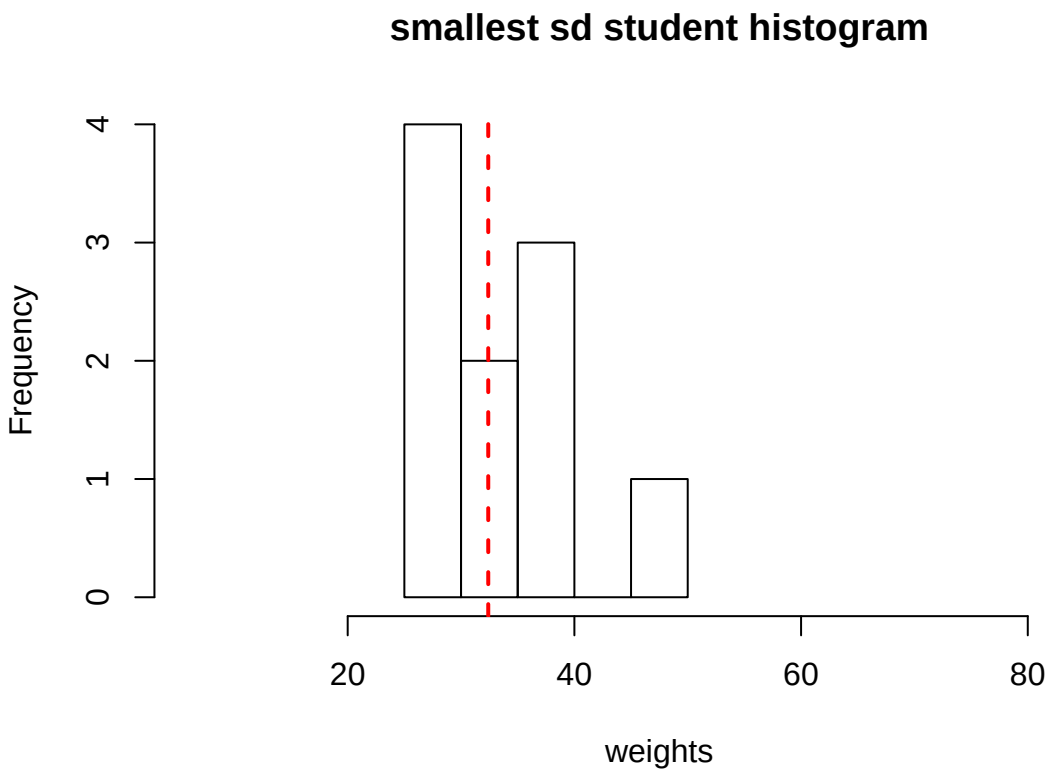
```
#smallest sd
h1<-(head(d1[with(d1,order((d1$SD),decreasing=FALSE)),],n=1))
h1[,c(1,12)]
```

```
##    students...1.      SD
## 14          7656 7.81736
```

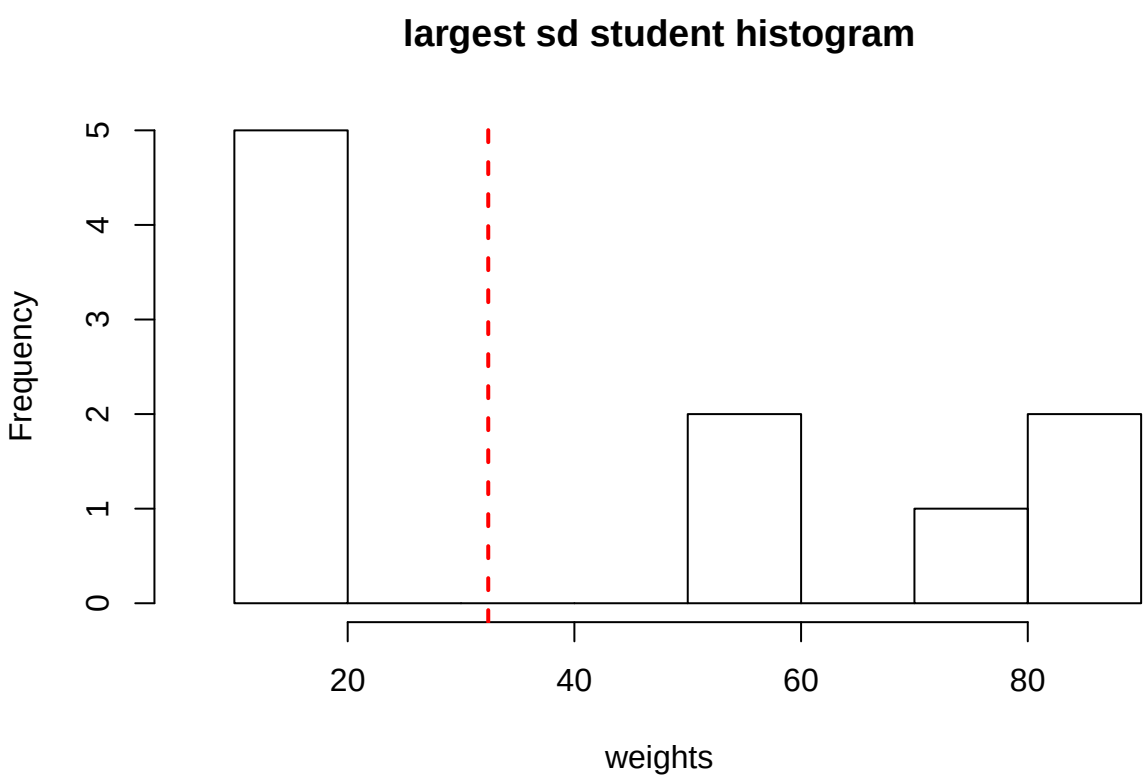
```
#highest sd
h2<-(head(d1[with(d1,order((d1$SD),decreasing=TRUE)),],n=1))
h2[,c(1,12)]
```

```
##    students...1.       SD
## 27          7582 31.09126
```

```
#h1[,c(-1,-12)]
hist(as.numeric(h1[,c(-1,-12)]),main="smallest sd student histogram",xlab="weights",
     xlim=extendrange(blocks$weight))
abline(v=mean(blocks$weight),col="red", lwd=2, lty=2)
```




```
hist(as.numeric(h2[,c(-1,-12)]),main="largest sd student histogram",xlab="weights",
     xlim=extendrange(blocks$weight))
abline(v=mean(blocks$weight),col="red", lwd=2, lty=2)
```



The student with the smallest standard deviation has chosen blocks whose weights measurements are close to one another, all in the range of 22 to . Since the mean of all blocks fall within the histogram, the student has chosen blocks whose mean is close to the actual mean. The student with the largest standard deviation has chosen blocks whose weights measurements are far from one another. Since the mean of all blocks fall where there is a break in the histogram, the mean of chosen blocks isnt that close to the actual value. It means the student has chosen blocks whose average is not close to the actual mean.

g. *(3 marks)* Comment on the quality of this judgment sampling plan, making reference to any of the results calculated above.

The judgement sampling plan is not good as the both sampling bias and sampling variability are not very low. The samples can vary a lot from student to student since it depends on the student as to what blocks they each select as per their judgement as that can be seen from the standard deviation values calculated. Also, most of the sample errors are much higher than zero as seen from the histogram.