EDA A5 Q4

Data

Set up the following:

```
## Set this up for your own directory
#imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
#dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
#path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
source(file = "/Users/rudranibhadra/Downloads/graphicalTests.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/generateData.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/numericalTests.R", echo = FALSE)
```

The full data set is then read in as:

```
labData <- read.csv("/Users/rudranibhadra/Downloads/labData.csv")
```

The data can be subsetted according to the three different experimental plans.

a. *(1 mark)* Select that subset of the data corresponding to the randomized plan. Assign it to the variable `randomized`. Show your code.

```
randomized<-labData[labData$type=='randomized',]
```

Analysis

b. *(4 marks)* Plot the (x, y) pairs from all of the randomized data

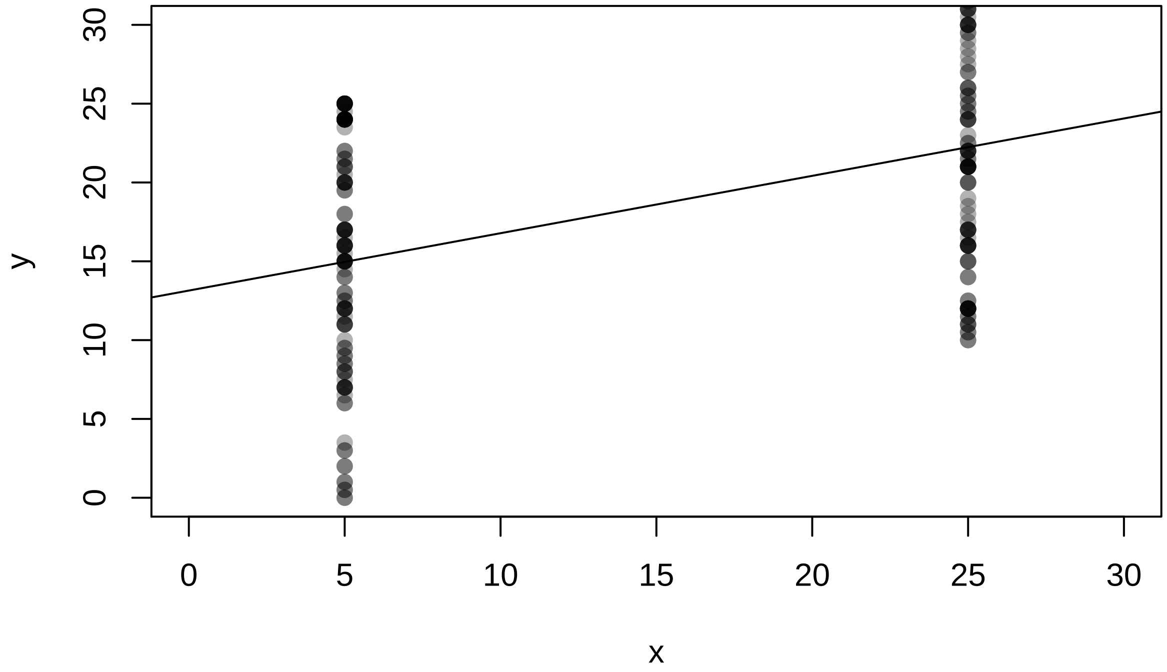
Use `xlim = c(0, 30), ylim = c(0,40), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

Label the plot meaningfully.

Fit a straight line model of y on x and add this fitted line to the plot. Save the fit object. Report the value of the slope estimate.

Show your code.

```
plot(randomized$x,randomized$y,xlim = c(0, 30), ylim = c(0,30), pch = 19,
     col = adjustcolor("black", 0.3),xlab='x',ylab='y' )
fit<-lm(randomized$y~randomized$x)
abline(fit)
```



```
#slope estimate
fit$coefficients['randomized$x']
```

```
## randomized$x
##    0.3638889
```

c. **Learning from repetition.** Each team executed the same plan. Moreover, each team replicated that execution. To gain a better appreciation of the qualities of that plan, we investigate the individual team estimates of β.

d. *(2 marks)* Separate the data into two subsets, one for each `rep`. Assign the two subsets to the variables `rand1` and `rand2` for replicates 1 and 2. Show your code.

```
rand1<-randomized[randomized$rep==1,]
rand2<-randomized[randomized$rep==2,]
```

ii. *(4 marks)* For each replication, fit a separate line for each team's data. For each replication, capture the slope estimates of each team's fit and collect these into a single vector. Call the vector for replication 1's slope estimates `slopes1` and the same for replication 2's `betas2`.

Show your code.

```
betas1<-c()
for(i in 1:18){
  r1<-rand1[rand1$team==i,]
  f1<-lm(r1$y~r1$x)
  betas1[i]<-(f1$coefficients['r1$x'])
}
#betas1
```

```r
betas2<-c()
for(i in 1:18){
  r2<-rand2[rand2$team==i,]
  f2<-lm(r2$y~r2$x)
  betas2[i]<-(f2$coefficients['r2$x'])
}
#betas2
```

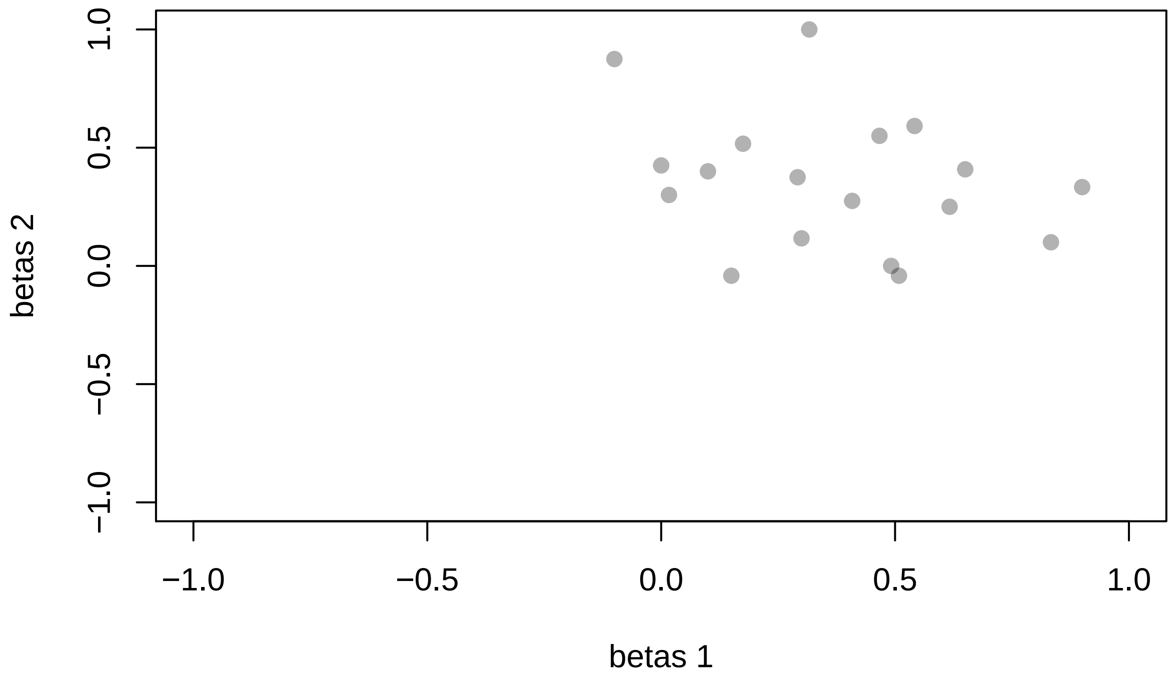
ii. *(4 marks)* Plot the $(betas1, betas2)$ pairs from the randomized data

Use `xlim = c(-1, 1), ylim = c(-1, 1), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

Label the plot meaningfully.

Show your code.

```r
plot(betas1,betas2,xlim = c(-1, 1), ylim = c(-1, 1), pch = 19,
     col = adjustcolor("black", 0.3),xlab='betas 1',ylab='betas 2')
```



iii. *(4 marks)* Test the hypothesis that the team paired slope estimators, $(\widetilde{\beta}_1, \widetilde{\beta}_2)$, based on replicates 1 and 2, are independently distributed. That is test $H_0 : \widetilde{\beta}_1 \perp\!\!\!\perp \widetilde{\beta}_2$.

Use `numericalTest()` with the appropriate choices of discrepancy measure and generation function.

Show your code.

Write up your conclusion about the independence.

```
h<-data.frame(betas1,betas2)
numericalTest(list(x=h$betas1,y=h$betas2),discrepancyFn=slopeDiscrepancy,generateFn =mixCoords)
```

```
## [1] 0.2305
```

Since the p value is greater than 0.05, it indicates weak evidence against the null hypothesis that betas1 and betas2 are independent and hence they are most likely independent.

iv. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams for **replicate 1** only.

Show your code.

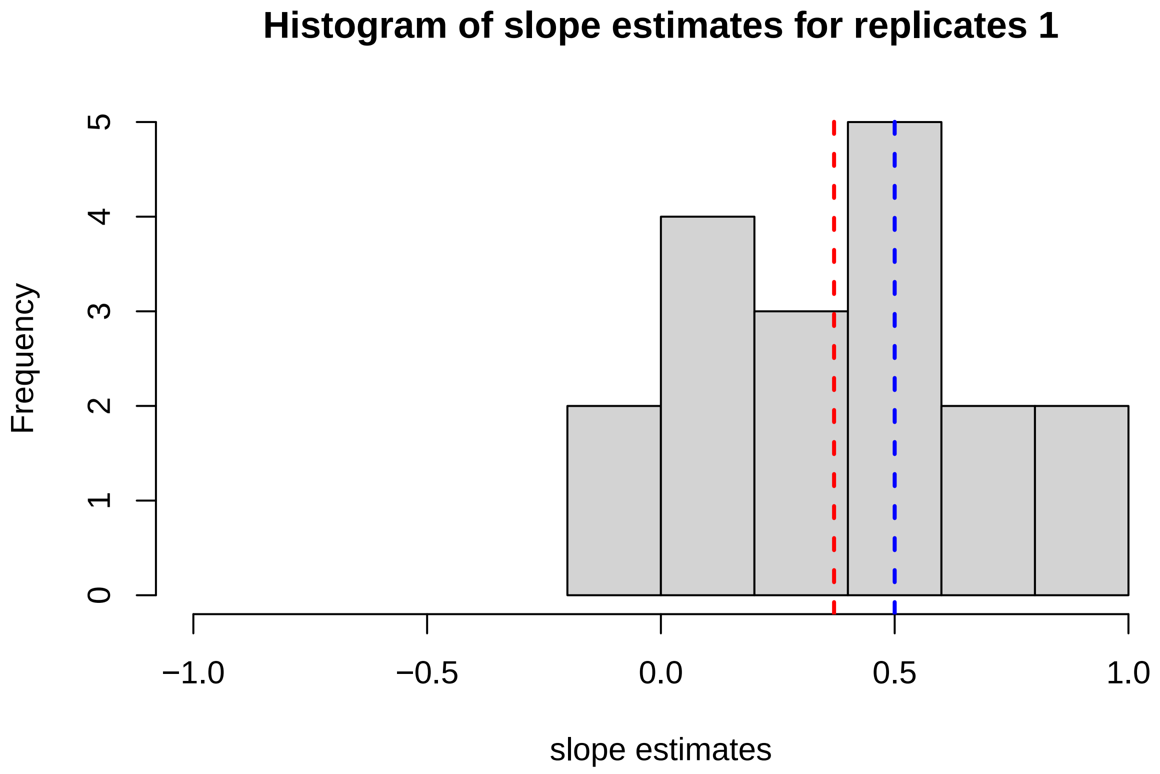
Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

Print the average and standard deviation of the slope estimates.

```
hist(betas1,xlim = c(-1, 1), col = "lightgrey",main = 'Histogram of slope estimates for replicates 1',
     xlab='slope estimates'  )
abline(v=mean(betas1),col='red',lwd=2,lty=2)
abline(v=0.5,col='blue',lwd=2,lty=2)
```



```
mean(betas1)
```

```
## [1] 0.3703704
```

```r
sd(betas1)
```

```
## [1] 0.283351
```

v. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams for **replicate 2** only.

Show your code.

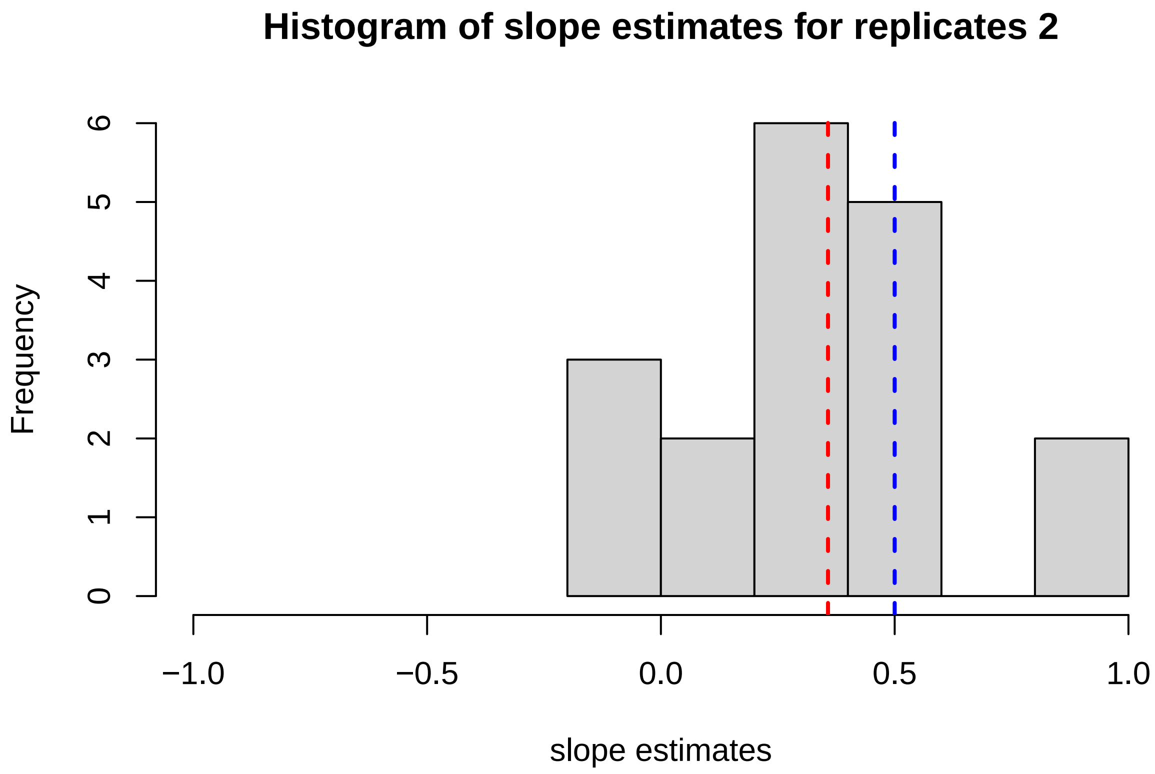
Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

Print the average and standard deviation of the slope estimates.

```r
hist(betas2,xlim = c(-1, 1), col = "lightgrey",main = 'Histogram of slope estimates for replicates 2',
     xlab='slope estimates')
abline(v=mean(betas2),col='red',lwd=2,lty=2)
abline(v=0.5,col='blue',lwd=2,lty=2)
```



```r
mean(betas2)
```

```
## [1] 0.3574074
```

```r
sd(betas2)
```

```
## [1] 0.2869847
```

vi. *(3 marks)* For all teams, draw a meaningfully labelled histogram of the average of the two individual slope coefficient estimates (over the two replicates).

Show your code.

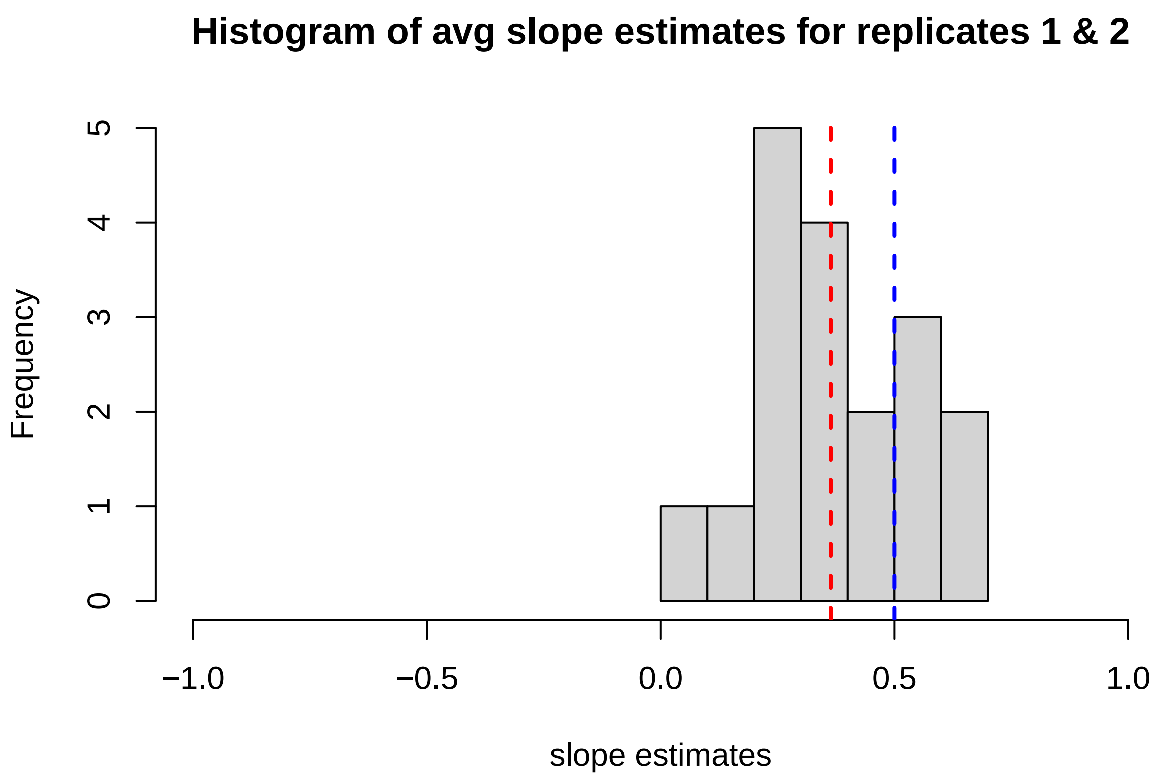
Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

Print the average and standard deviation of the slope estimates.

```r
s<-rowMeans(cbind(betas1,betas2))
hist(s,xlim = c(-1, 1), col = "lightgrey",main = 'Histogram of avg slope estimates for replicates 1 & 2
     xlab='slope estimates' )
abline(v=mean(s),col='red',lwd=2,lty=2)
abline(v=0.5,col='blue',lwd=2,lty=2)
```



```r
mean(s)
```

```
## [1] 0.3638889
```

```r
sd(s)
```

```
## [1] 0.1692267
```

Conclusion

e. *(3 marks)* What do you conclude about the quality of team slope estimates from the randomized study?

The total average slope estimates from each subset (replicate) as well the average of both replicates is close to the the true slope estimate. This shows that the quality of slope estimates from the randomized study is fairly good.

f. *(2 marks)* What do you conclude about the value of having each team average their replicates from the randomized study?

By having each team average their replicates, we can see that the average slope estimate (0.363) is very close to the true slope estimate, although it is slightly further away from the true average than the average of slopes for replicates 1 (0.37) but closer than the average of slopes for replicates 2 (0.357).

g. *(2 marks)* What effect, if any, has been produced by a lurking variable? Explain.

Here we are analysing the relationship between x and y taking type as randomized and taking rep as 1 and 2. Here a lurking variable does not have much effect on the relationship of x and y since it is a positive relationship.