# POLS 602 PS2

Julia Reyes

Setup

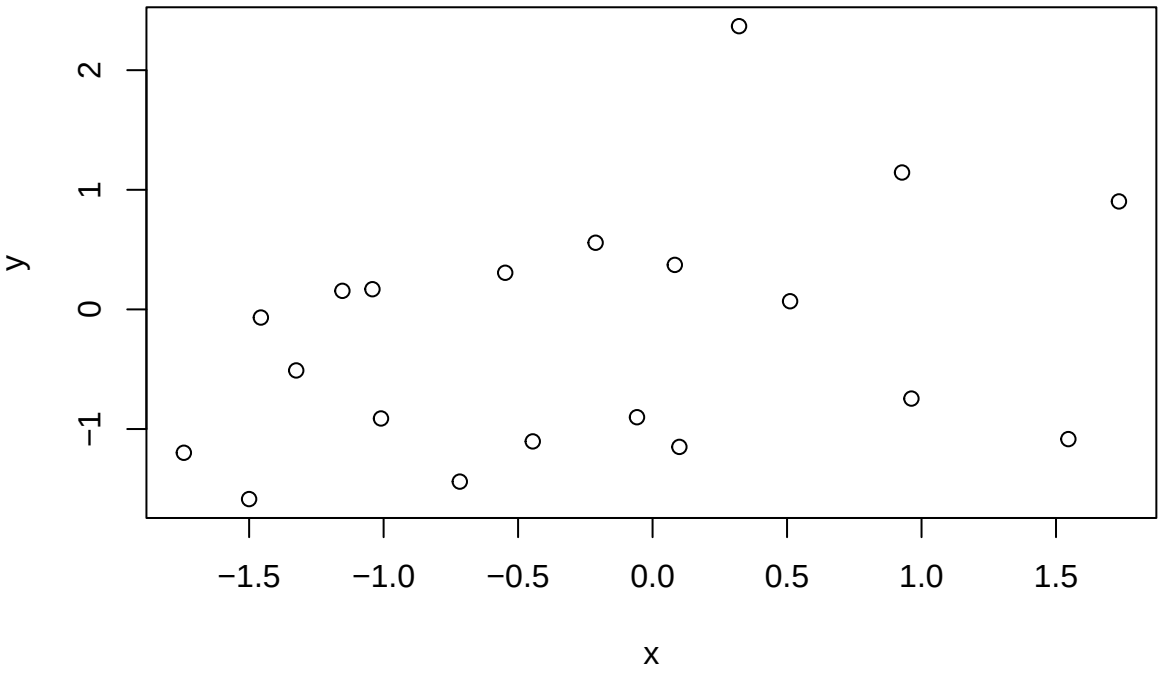
```
set.seed(25)
```

Use rnorm to create two random variables with 20 observations each

```
x <- rnorm(20)
y <- rnorm(20)
```

Find correlation between x and y

```
plot(x, y) #just to visualize
```




```
cor(x,y)
```

```
## [1] 0.3488868
```

Repeat process many times

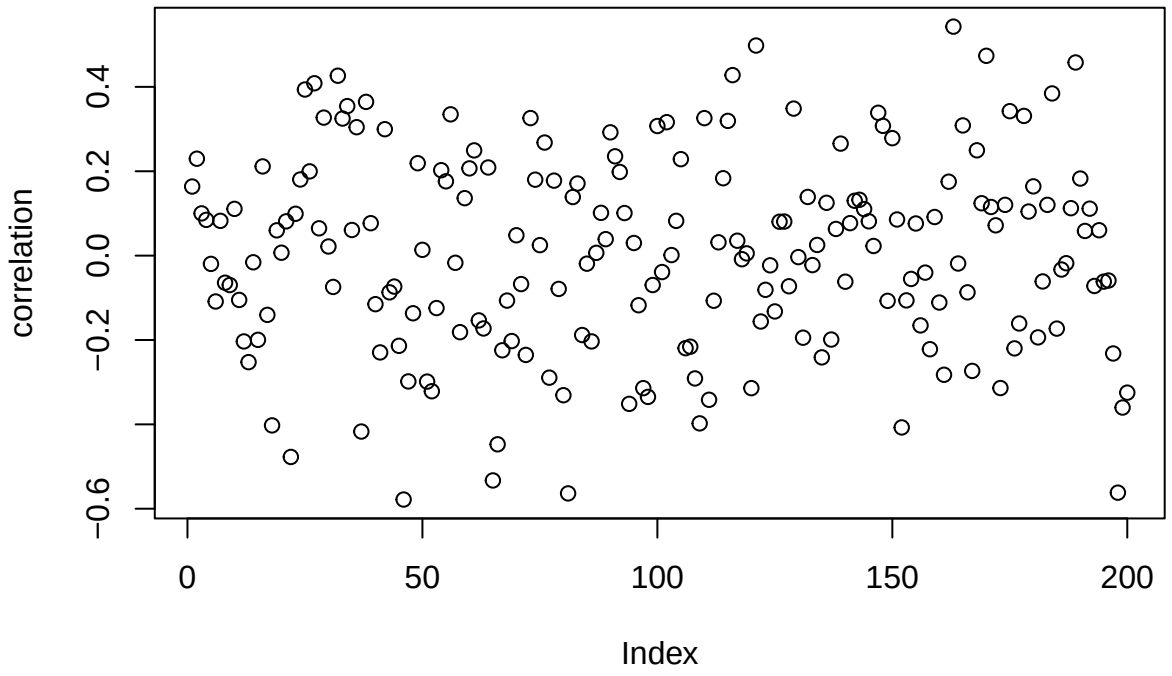
```
#create empty container for correlation coefficients
correlation <- numeric(200)

#repeat using for loop
for(n in 1:200){
    x <- rnorm(20)
    y <- rnorm(20)
```

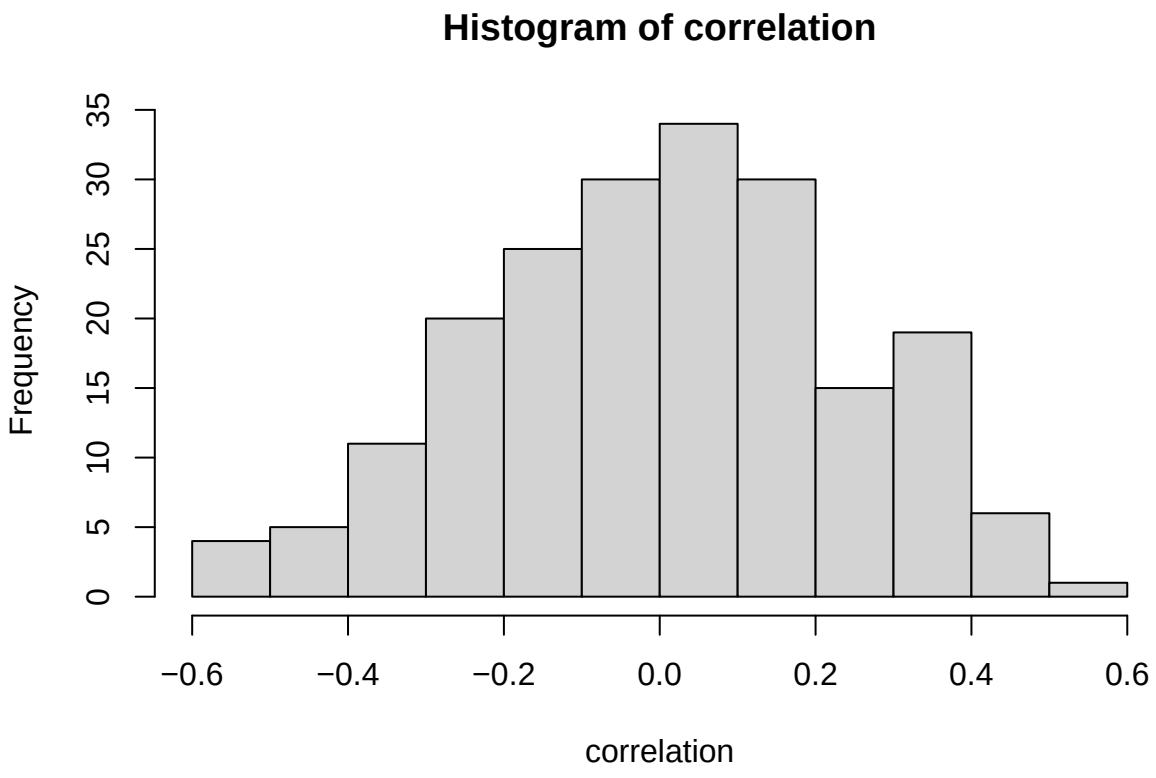
```
    correlation[n] <- cor(x,y)
}
```

Plot correlation coefficients

```
#decided to plot and also create a histogram to get a better sense of the spread
plot(correlation)
```




```
hist(correlation)
```



Calculate standard deviation

```
sd(correlation)
```

```
## [1] 0.23081
```

On average, what would we expect the correlation between the two variables to be? What does this distribution tell us about sample estimates of population parameters?

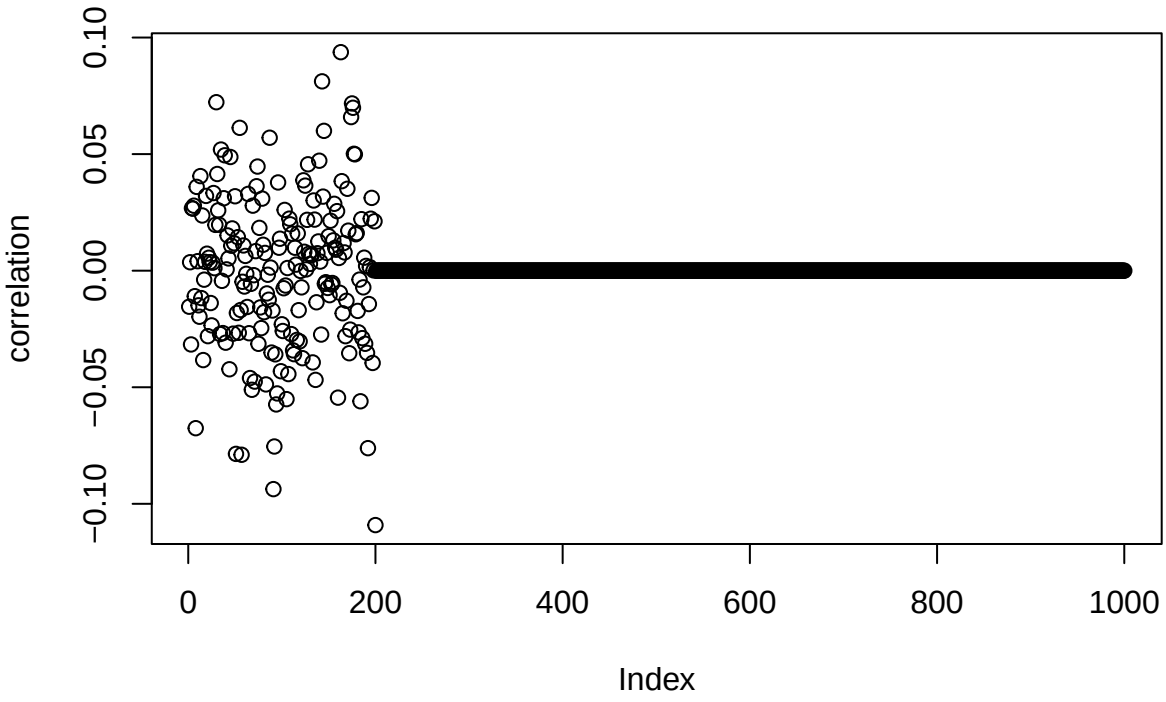
On average, we would expect the correlation to be zero, despite some noise that might not lead to a perfectly zero correlation for any given observation. This tells us that we can expect an average of zero correlation between variables which are not related to one another. This also tells us that if we have a small enough sample size, we might infer that there is a correlation between variables when there actually is not one, because of the random noise. If we increase the sample size, it becomes more clear that the correlation is actually zero.

Repeat with sample size of 1000

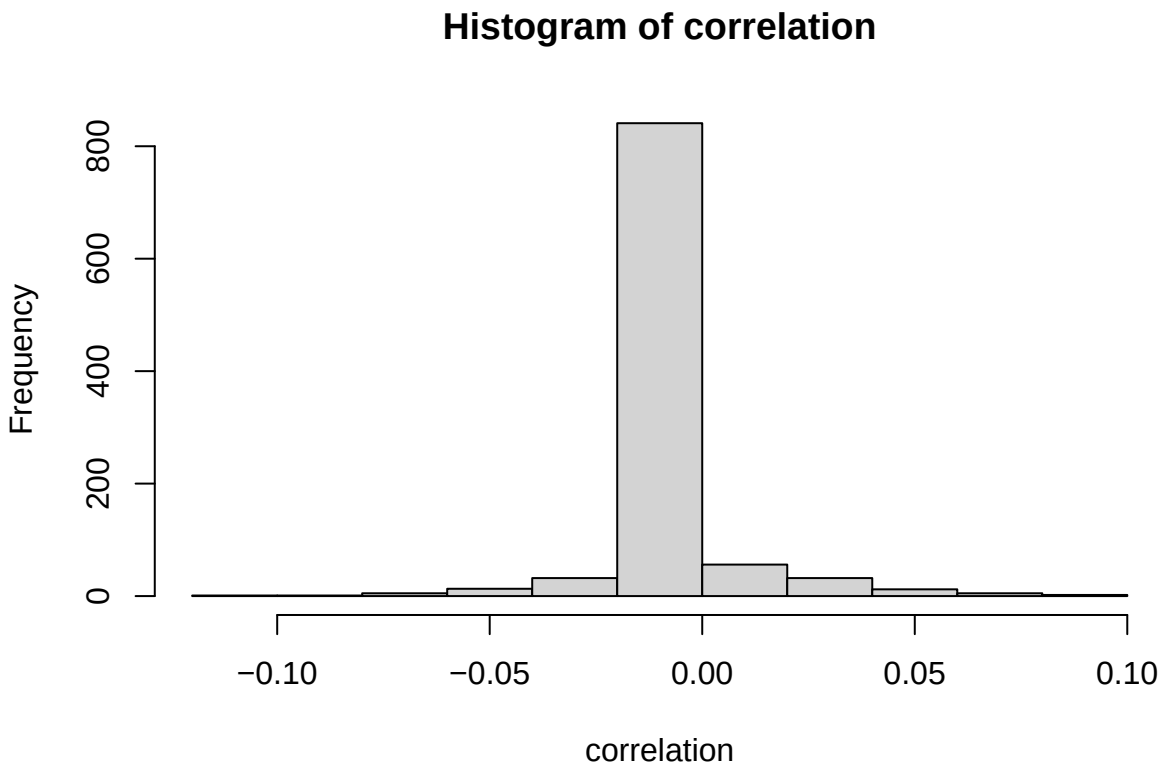
```
#create empty container for correlation coefficients
correlation <- numeric(1000)

#repeat using for loop
for(n in 1:200){
    x <- rnorm(1000)
    y <- rnorm(1000)
    correlation[n] <- cor(x,y)
}

#plot correlation coefficients
plot(correlation)
```



```r
hist(correlation)
```




The results for the sample size of 1000 differ because the correlation coefficients are more clustered around zero in the histogram. This means that if we increase our number of observations of the correlation between two unrelated, random variables, we can more obviously see that they have zero correlation, despite some minor remaining random noise.

Create three random variables X, Y, and Z with the causal relationships stated

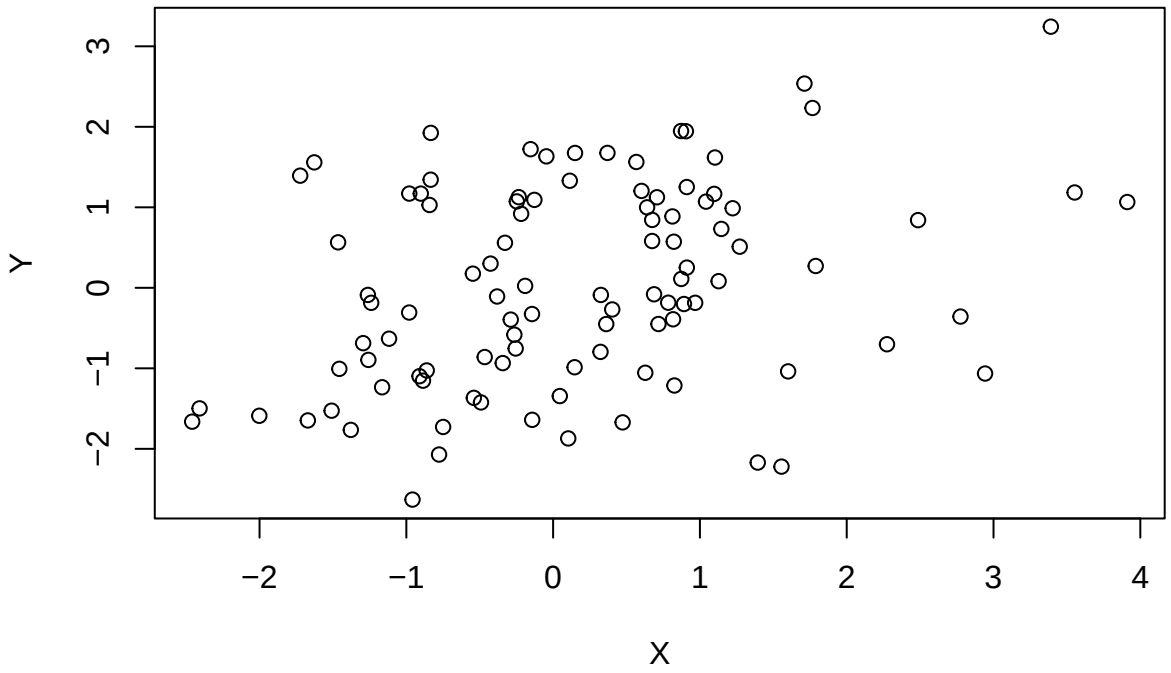
```r
#create random variable Z

Z <- rnorm(100)

#create X and Y as some function of Z

X <- Z + rnorm(100)
Y <- Z + rnorm(100)

#plot X and Y on scatterplot
plot(X, Y)
```




```
#find correlation between X and Y
cor(X, Y)
```

```
## [1] 0.3355625
```

This tells us that, because X and Y are both causally related to the same random variable Z, we should expect that they will have some correlation to one another. They are not causally related to one another, only associated (unlike the previous question in which both variables were independent and not associated).