Problem Set 2

Katherine Jones

2025-10-22

#Simulation Question 1

Create two random variables with 20 observations each and calculate the correlation between the two variables.

The two random variables created were variable K and variable J which each have 20 observations. The correlation between the two variables is -0.22.

```
#rnorm() function for two random variables
set.seed(1)
K <- rnorm(20)
J <- rnorm(20)
#calculate correlation
cor(K, J)
```

```
## [1] -0.2175249
```

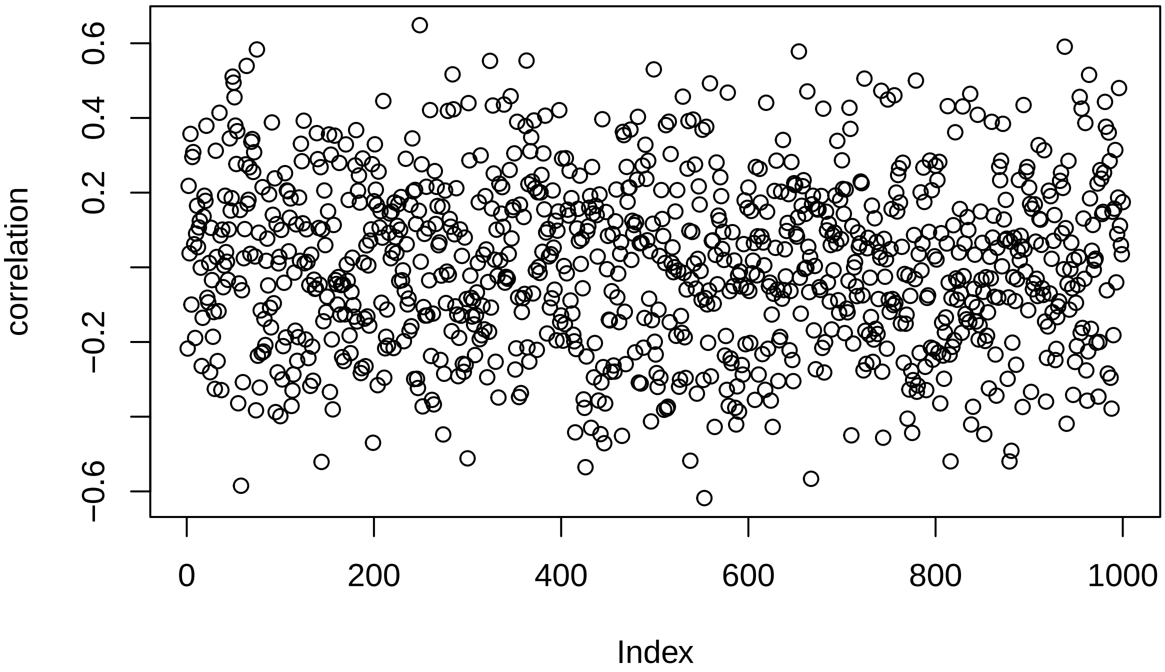
Repeat the process many times

```
set.seed(1)

correlation <- numeric(1000) #container

for (i in 1:1000) { #the repeating part
  K <- rnorm(20)
  J <- rnorm(20)
  correlation[i] <- cor(K, J)
}
```

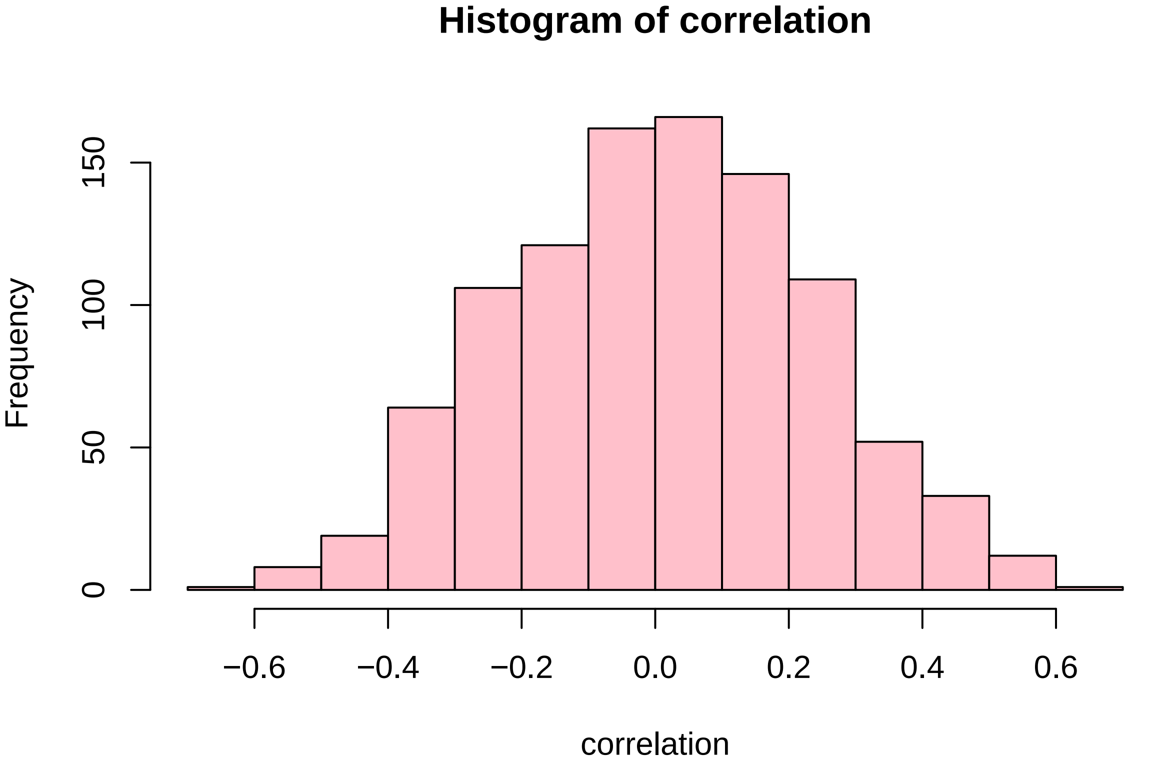
Plot the distribution of the correlation coefficients and report the standard deviation

The standard deviation of the correlation coefficients from zero is 0.23.

```
plot(correlation)
```



```r
hist(correlation, col = "pink") #easier to visualize
```



```r
sd(correlation) #standard deviation
```

```
## [1] 0.2267199
```

On average, we would expect that the correlation between the two variable be 0 since these are two random variables with no relation to one another. This distribution shows that even for random variables within a sample estimate of population parameters will still have some random noise, especially when the population parameters is a smaller measure.

#Simulation Question 2

Repeat the steps in Question 1

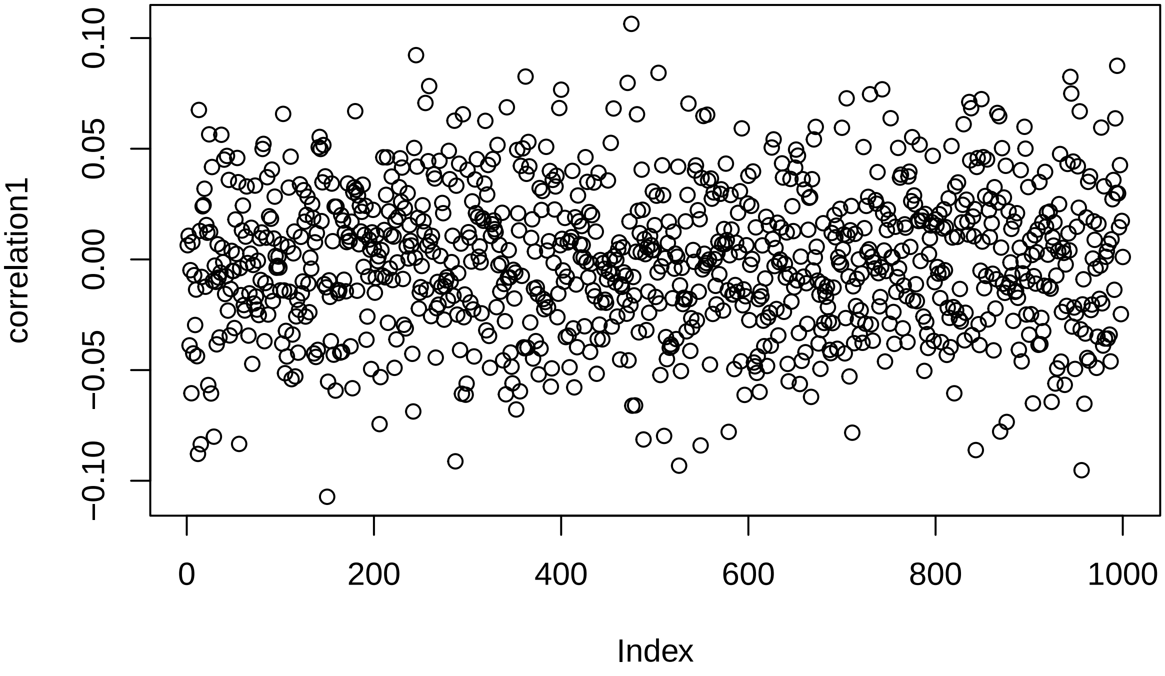
When the random variable's observation size was increased, the correlation between the two shrunk to 0.006.

```
#used new variable letters as to not confuse myself but functionally they are the same
set.seed(1)
L <- rnorm(1000)
M <- rnorm(1000)
#calculate correlation
cor(L, M)
```

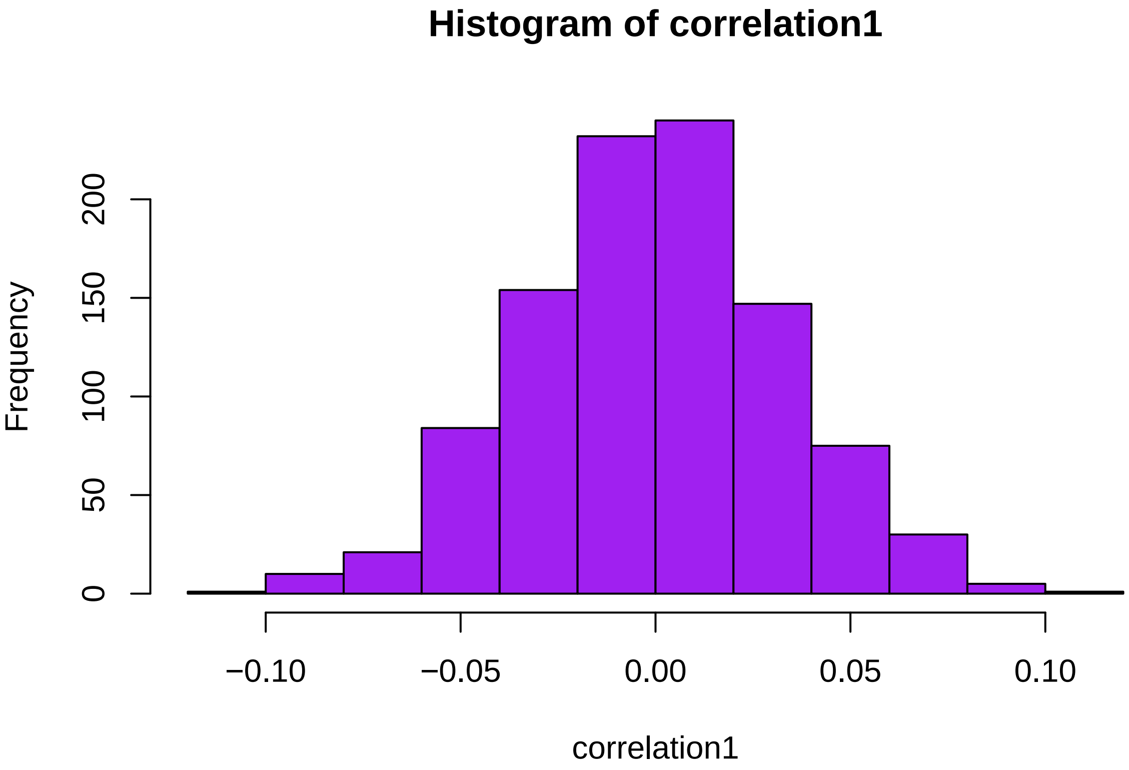
```
## [1] 0.006401211
```

Repetition stage

```
set.seed(1)

correlation1 <- numeric(1000) #new container for new question

for(i in 1:1000){ #same for loop as above
  L <- rnorm(1000)
  M <- rnorm(1000)
  correlation1[i] <- cor(L, M)
}

plot(correlation1) #scatterplot
```



```
hist(correlation1, col = "purple") #histogram
```



```
sd(correlation1) #standard deviation
```

```
## [1] 0.03276884
```

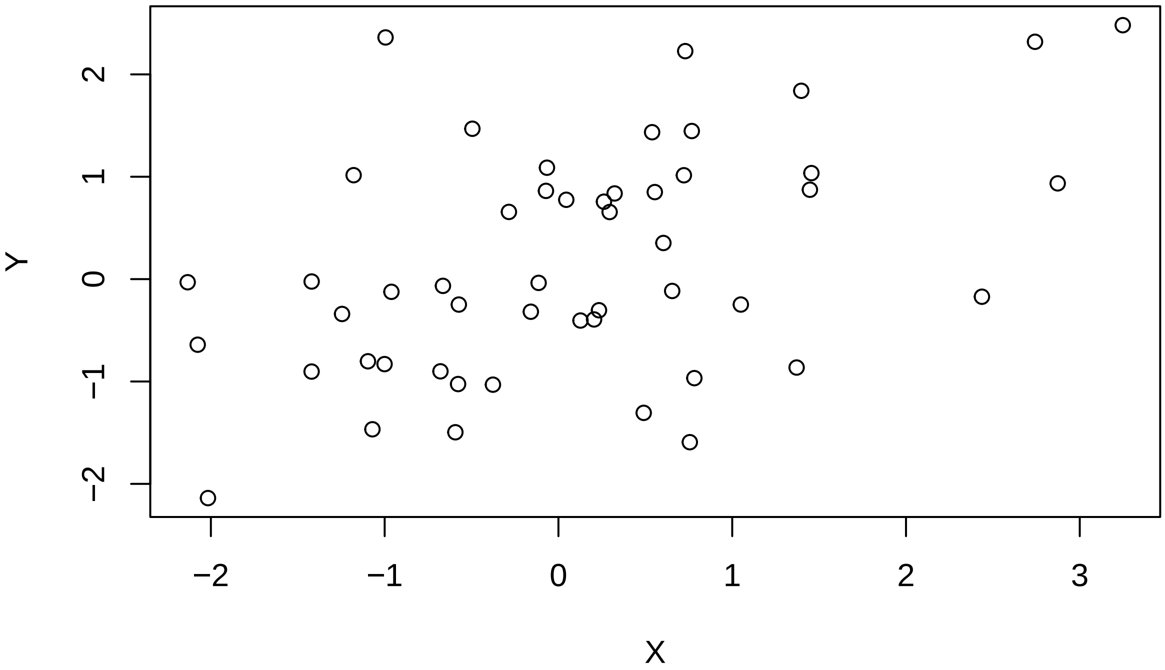
Compared to the rnorm() simulation with 20 observations, the correlation for the rnorm() simulation with 1000 observations is closer to 0. This is an expected difference in the results because as we increased our observation size of the two random variables, the distribution of the correlation coefficients will get smaller, and eventually reaching 0, as our random sample observations grows to resemble the population parameters better. This difference is also observed in the calculation of the standard deviation for each result. For the simulation with 20 observations, the standard deviation for correlation coefficients is 0.23, while for the simulation with 1000 observations, the standard deviation is 0.03. This difference in the two samples shows that as our observation size grows, the random noise will be reduced and more representative of what we expect the correlation of two random variables to be. These results can also be seen visually in the scatterplots and histograms for each simulation where for the simulation with 1000 observations, the data points are more centered around 0 than in the simulation with 20 observations.

#Simulation Question 3 Create three random variables with a DAG relationship

```
Z <- rnorm(50) #causal variable
X <- Z + rnorm(50) #random variable 1
Y <- Z + rnorm(50) #random variable 2
```

Plot X and Y on a scatter plot and report their correlation

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



```
cor(X,Y)
```

```
## [1] 0.4674704
```

For variables X and Y which have no causal relationship, their correlation is 0.47. This tells us when we are interpreting correlations that correlation does not equal causation. For example, in this correlation simulation, we cannot say that X causes Y despite the high correlation because there is no direct causal relationship. If we were testing the correlation between Z and Y, then we could say that Z cause Y because we have that direct causal link. However, this simulation testing the correlation between X and Y shows that we cannot always interpret the correlation between two random variables to equal causation.