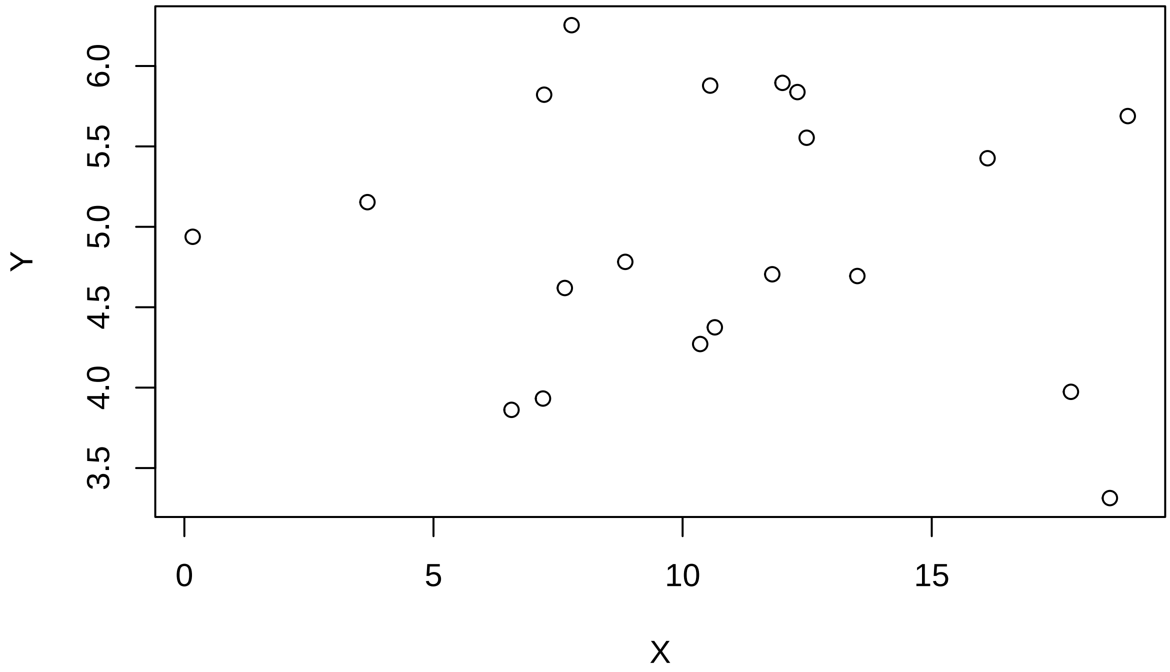
Problem Set 2

Zhengyu Xiao

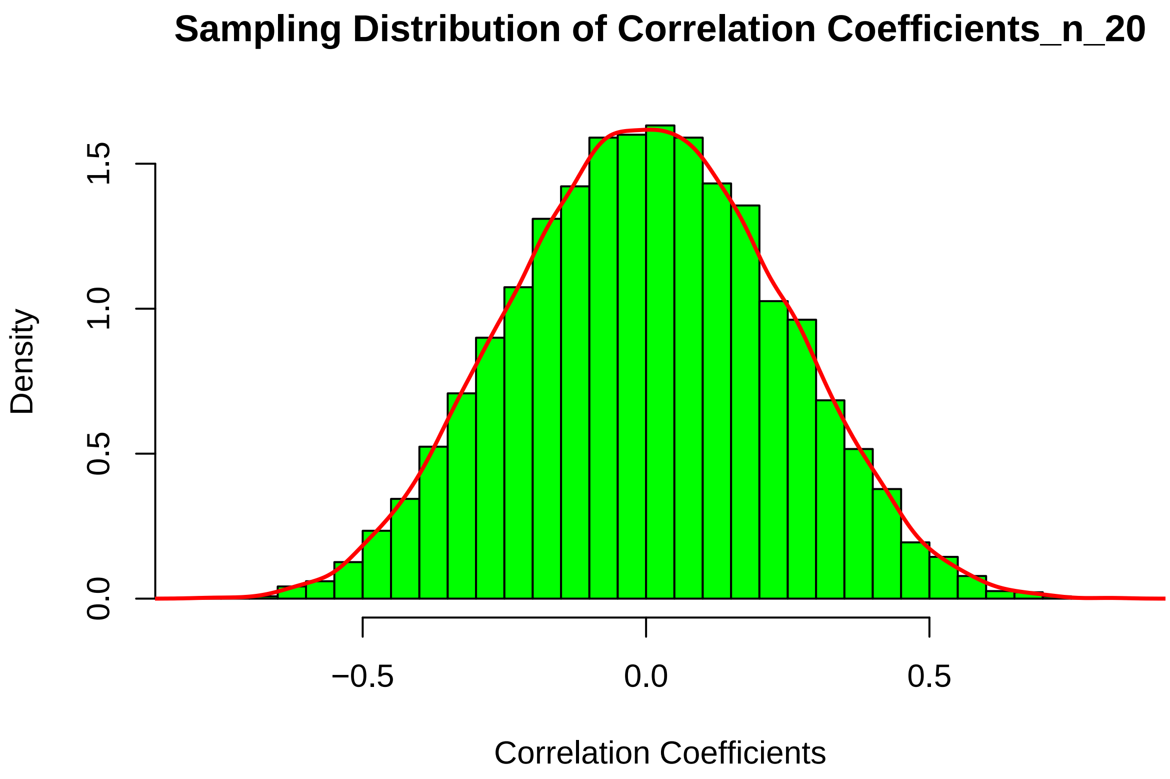
2025-10-19

1. Use the rnorm() function to create two random variables in R with 20 observations each. Then, calculate the correlation between the two variables. Repeat this process many times. Plot the distribution of the correlation coefficients and report the standard deviation. 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?

```
set.seed (123)

num_simulation <- 10000

Correlation <- numeric (num_simulation)

X <- rnorm (20, 10, 5)
Y <- rnorm (20, 5, 1)

correlation_1 <- cor(X, Y)

plot(X,Y)
```



```r
for (i in 1:num_simulation){

  X <- rnorm (20, 10, 5)
  Y <- rnorm (20, 5, 1)
  Correlation[i] <- cor (X,Y)
}

hist (Correlation,
      main = "Sampling Distribution of Correlation Coefficients_n_20",
      xlab = "Correlation Coefficients",
      breaks = 35,
      col = "green",
      freq = FALSE)
lines(density(Correlation), col = "red", lwd = 2)
```



```
print(sd(Correlation))
```

```
## [1] 0.2322784
```

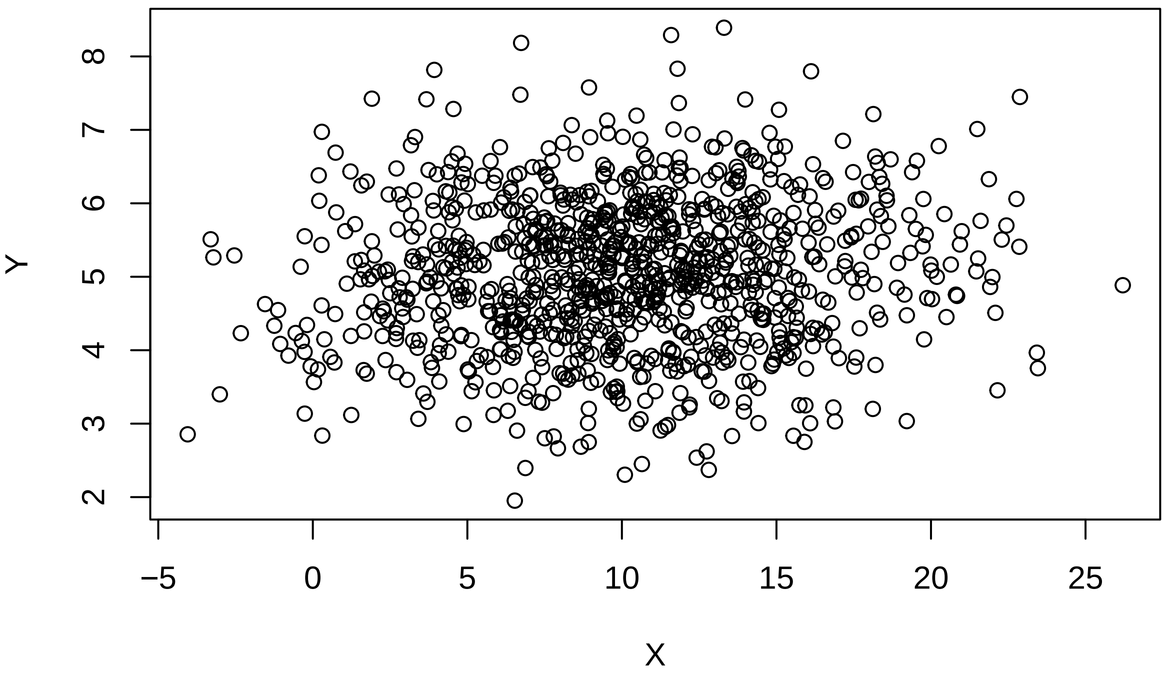
The distribution of correlation coefficients is plotted as above, the standard deviation is approximately 0.23. On average, we would expect the correlation between the two variables to be 0.

This distribution tells us that: There is a significant amount of sample variabilities, we have two totally uncorrelated variables and after repeating a large amount of calculation of coefficients, we have a range of correlation coefficients from approximately -0.5 to 0.5 with a standard error of around 0.23 while the population correlation coefficients is supposed to be zero. This means we cannot just repy on any single sample estimates to acquire population parameters.

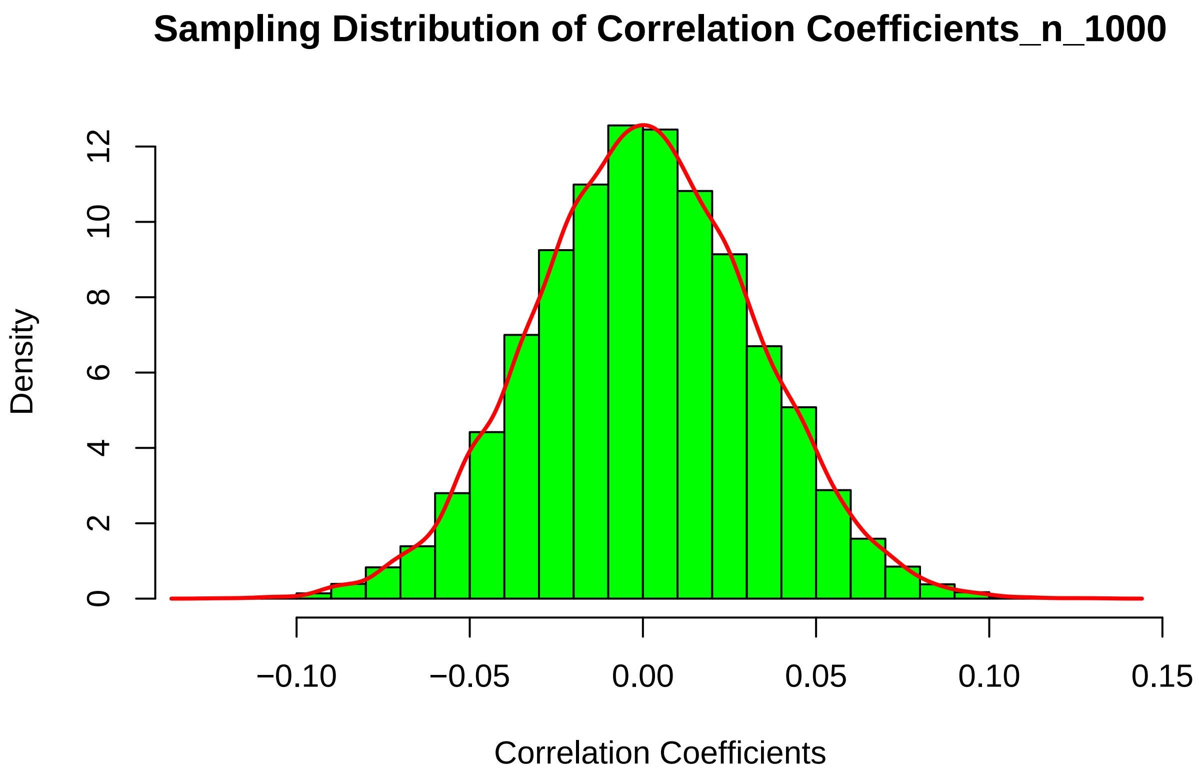
2. Repeat the previous step with a sample size of 1,000 and provide a substantive interpretation of how the results differ.

```
set.seed (123)

num_simulation <- 10000

Correlation <- numeric (num_simulation)

X <- rnorm (1000, 10, 5)
Y <- rnorm (1000, 5, 1)

correlation_1 <- cor(X, Y)
```

```
plot(X,Y)
```



```
for (i in 1:num_simulation){

  X <- rnorm (1000, 10, 5)
  Y <- rnorm (1000, 5, 1)
  Correlation[i] <- cor (X,Y)
}

hist (Correlation,
      main = "Sampling Distribution of Correlation Coefficients_n_1000",
      xlab = "Correlation Coefficients",
      breaks = 35,
      col = "green",
      freq = FALSE)
lines(density(Correlation), col = "red", lwd = 2)
```



```
print(sd(Correlation))
```

```
## [1] 0.03191016
```

From the comparing results with a sample size of 1000 and a sample size of 20, we can interpret that:

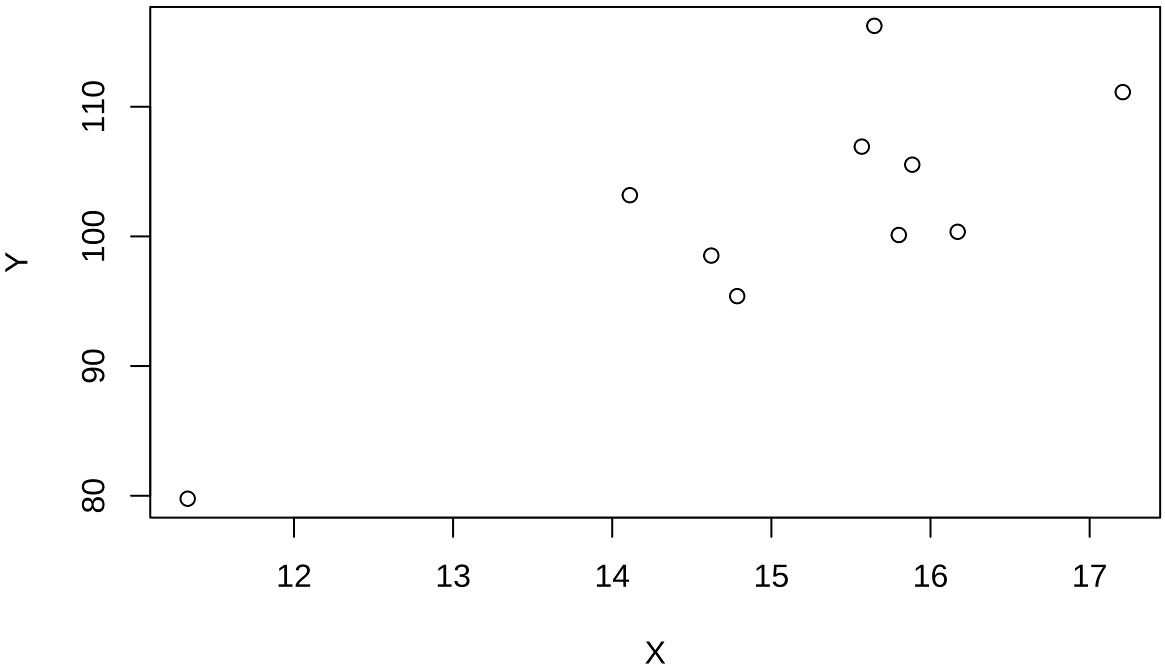
As sample size increases, the distribution of correlation coefficients is narrower and the standard deviation decreases, although the population correlation is still zero. This means that with a larger sample size in any random single calculation of correlation coefficients, we have a reduction in sampling errors and more reliable estimate of sample parameters because we covered more traits to create a more similar sample statistics compared to the population statistics according to the Law of Large Numbers.

3. Create three random variables in R that have the following causal relationship: Z causes both X and Y, but X and Y have no causal relationship. Start by generating Z as a random variable, then create X and Y as some function of Z plus random noise. Plot X and Y on a scatter plot and report their correlation. What does this tell us about interpreting correlations?

a) Single experiment with samples of 10

```
Z <- rnorm (10, 10, 1)
epsilon_ZX <- rnorm (10, 0, 1)
```

```
epsilon_ZY <- rnorm (10, 0, 2)
X <- 1.5*Z + epsilon_ZX
Y <- 10 * Z + epsilon_ZY

plot(X,Y)
```

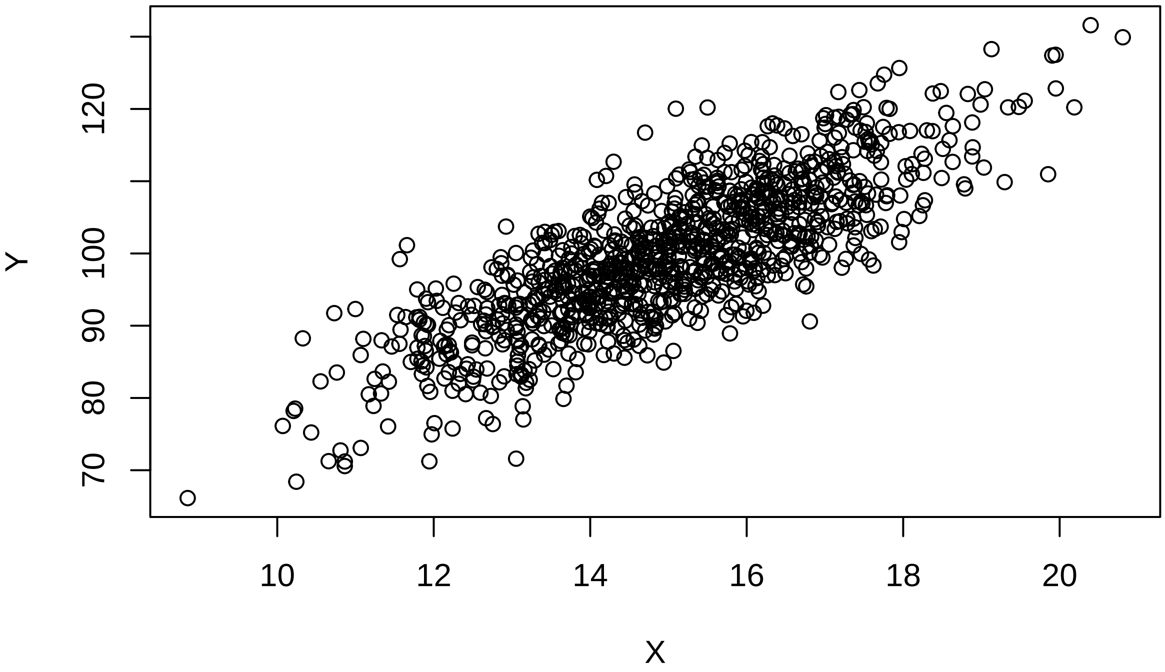


```
cor(X,Y)
```

```
## [1] 0.8163363
```

b) Single experiment with samples of 1000

```
Z <- rnorm (1000, 10, 1)
epsilon_ZX <- rnorm (1000, 0, 1)
epsilon_ZY <- rnorm (1000, 0, 2)
X <- 1.5*Z + epsilon_ZX
Y <- 10 * Z + epsilon_ZY

plot(X,Y)
```

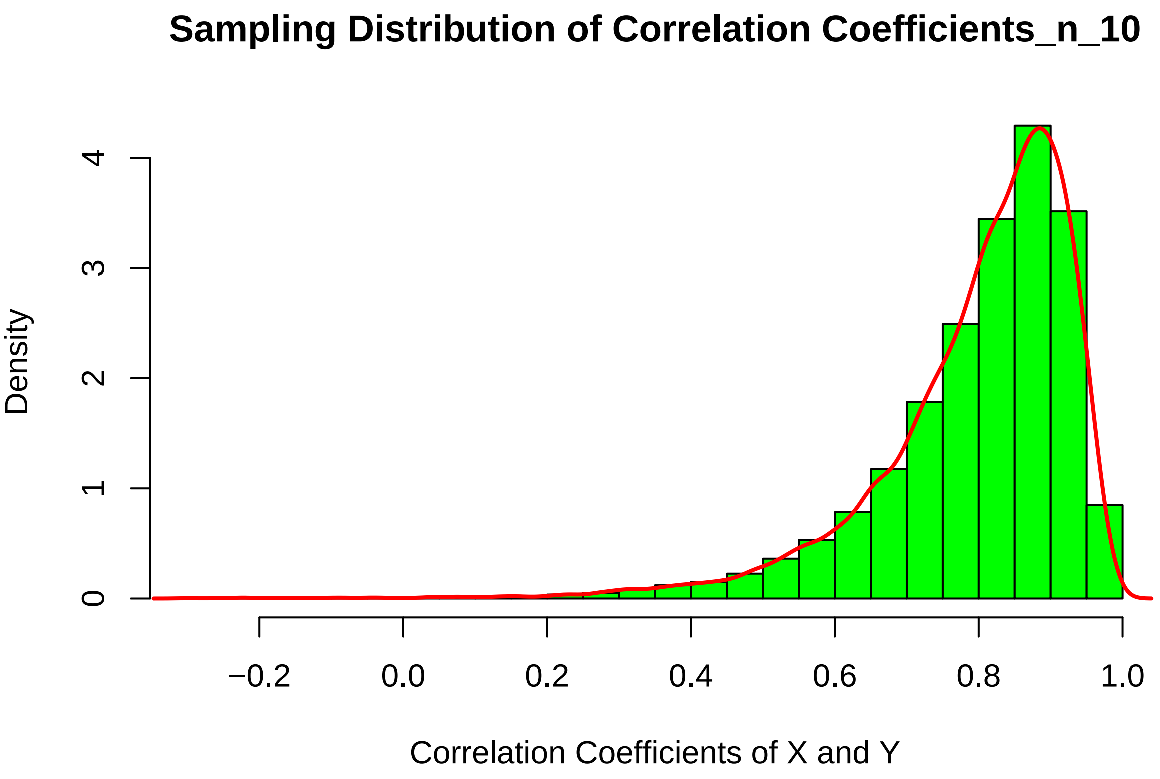


```
cor(X,Y)
```

```
## [1] 0.8100037
```

c) Repeat 10000 times with samples of 10

```
set.seed (123)

num_simulation <- 10000

Correlation <- numeric (num_simulation)

for (i in 1:num_simulation){

  Z <- rnorm (10, 10, 1)
  epsilon_ZX <- rnorm (10, 0, 1)
  epsilon_ZY <- rnorm (10, 0, 2)
  X <- 1.5*Z + epsilon_ZX
  Y <- 10 * Z + epsilon_ZY
  Correlation[i] <- cor (X,Y)
}

hist (Correlation,
      main = "Sampling Distribution of Correlation Coefficients_n_10",
      xlab = "Correlation Coefficients of X and Y",
      breaks = 35,
      col = "green",
      freq = FALSE)
lines(density(Correlation), col = "red", lwd = 2)
```



```
print(sd(Correlation))
```

```
## [1] 0.1377474
```

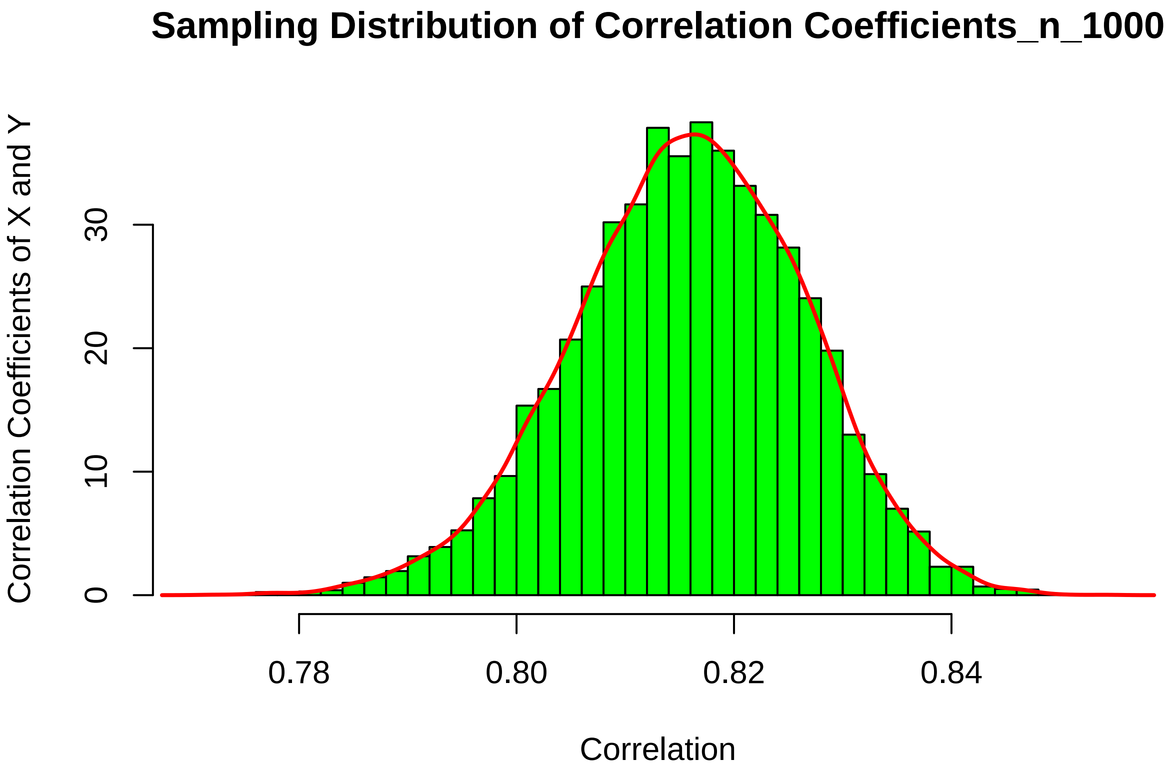
```
cor(X,Y)
```

```
## [1] 0.8596762
```

d) Repeat 10000 times with samples of 1000

```
set.seed (123)

num_simulation <- 10000

Correlation <- numeric (num_simulation)

for (i in 1:num_simulation){

  Z <- rnorm (1000, 10, 1)
  epsilon_ZX <- rnorm (1000, 0, 1)
  epsilon_ZY <- rnorm (1000, 0, 2)
  X <- 1.5*Z + epsilon_ZX
  Y <- 10 * Z + epsilon_ZY
  Correlation[i] <- cor (X,Y)
}
```

```r
hist (Correlation,
      main = "Sampling Distribution of Correlation Coefficients_n_1000",
      ylab = "Correlation Coefficients of X and Y",
      breaks = 35,
      col = "green",
      freq = FALSE)
lines(density(Correlation), col = "red", lwd = 2)
```



```r
print(sd(Correlation))
```

```
## [1] 0.01061732
```

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

```
## [1] 0.7984061
```

After running single and repeated experiments (10 and 1000 samples), we observe that X and Y are highly correlated even though there is no direct causal link between them. This demonstrates that correlation does not imply causation: two variables may be strongly correlated simply because they are both influenced by a third variable (Z). Thus, interpreting correlations without considering confounding factors can be misleading.