Problem Set 2

Neeraj Tom Savio

2025-10-22

```
## Introductory Code and Setting Up

rm(list = ls()) #Clear the environment
setwd("~/Desktop/POLS 602/R Files") #Designating the Working Directory
```

Simulation

Question 1

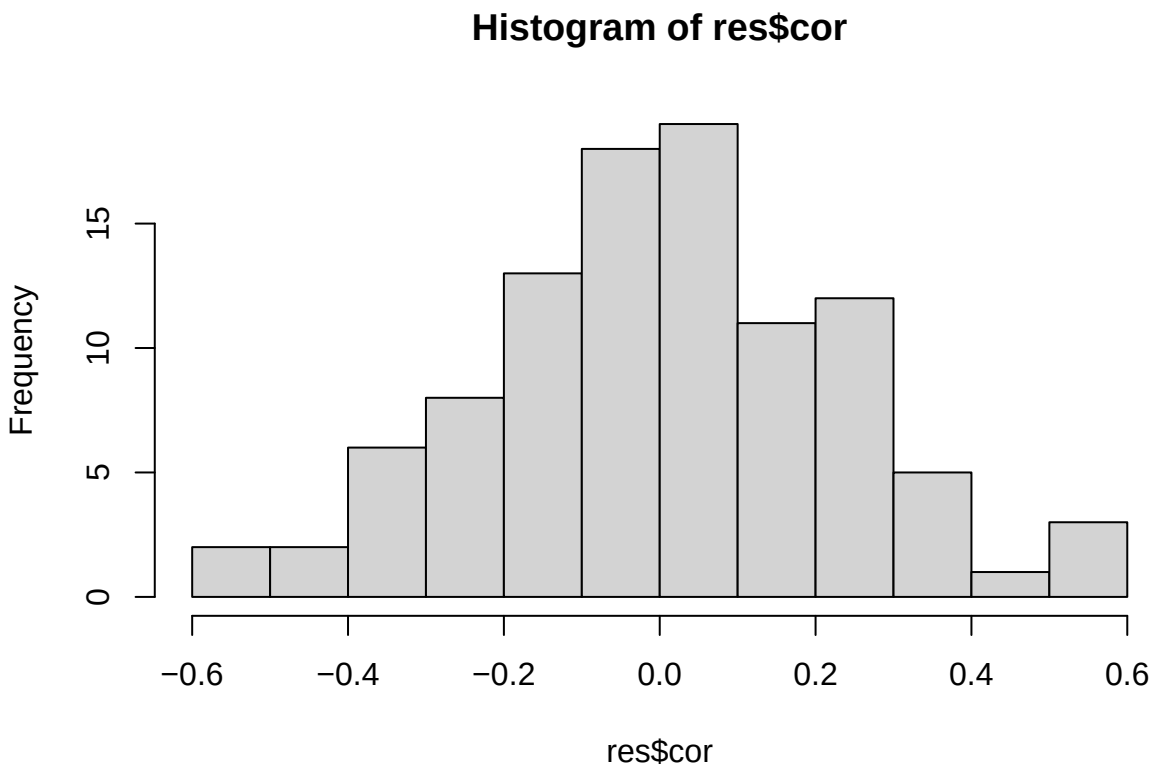
```
set.seed(1002003) #For reproductibility

res <- data.frame(cor = numeric()) #Creating an empty data frame

#Using a FOR loop to repeat the task 100 times
for(i in 1:100){
  a <- rnorm(20, mean = 20, sd = 5)
  b <- rnorm(20, mean = 33, sd = 7)
  cor_val <- cor(a, b)

  res[i, "cor"] <- cor_val #Saving the results to the empty data frame
}

hist(res$cor) #Creating a histogram to depict the frequency distributio
```




```
avg_cor <- mean(res$cor)
stan_dev <- sd(res$cor)

avg_cor
```

```
## [1] -0.002123023
```

```
stan_dev
```

```
## [1] 0.2307352
```

The average correlation between a and b is -0.002123023 and the standard deviation is 0.2307352.

We know that a and b are distinct random variables and their correlation should be 0. However, we can infer from the above distribution that sample estimates of population parameters vary. If we had run a single correlation, we could have gotten an erroneous chance correlation. However, repeating the same statistical exercise multiple times helps us to see how the sample estimates vary and get an average value that is much more likely to be closer to the population parameter.

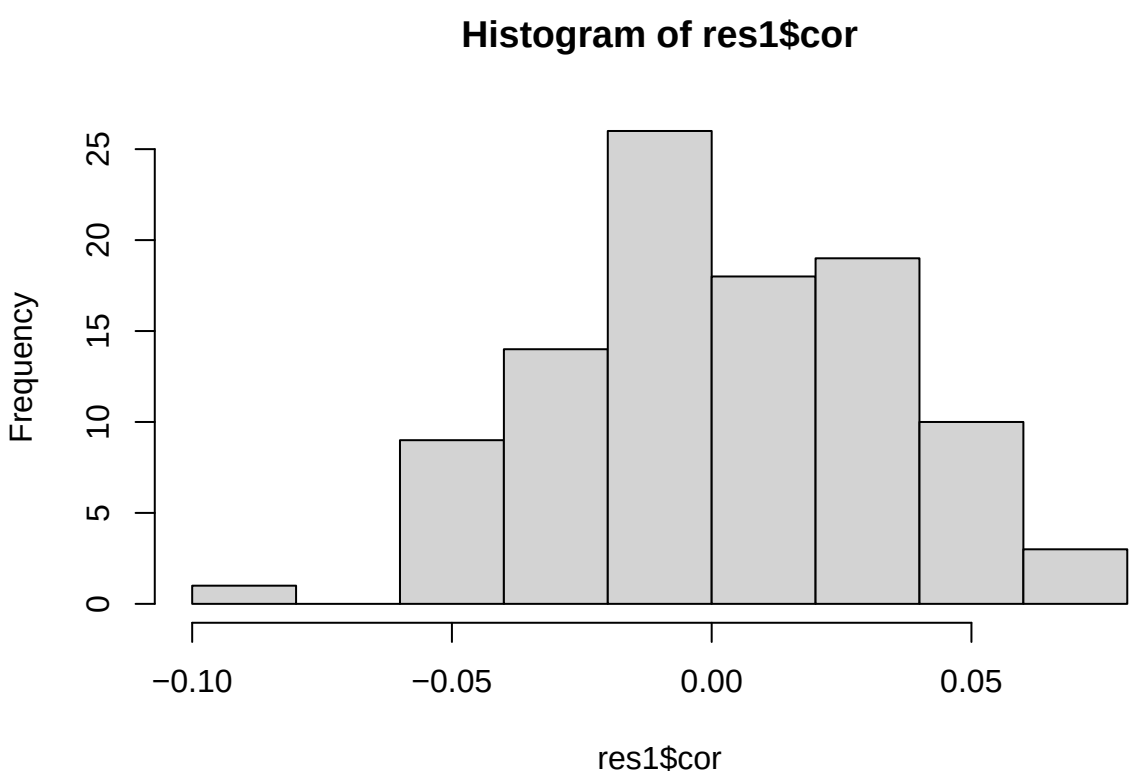
Question 2

```
## Repeating the process but with 1000 observations

res1 <- data.frame(cor = numeric())

for(i in 1:100){
  a <- rnorm(1000, mean = 20, sd = 5) #Pulling 1000 observations instead of 20
  b <- rnorm(1000, mean = 33, sd = 7)
  cor_val2 <- cor(a, b)
  res1[i, "cor"] <- cor_val2
}
```

```r
hist(res1$cor) #Creating a histogram of the new correlations
```




```r
avg_cor2 <- mean(res1$cor) #Average Correlation
stan_dev2 <- sd(res1$cor) #Standard Deviation of the correlations

avg_cor2
```

```
## [1] 0.003448129
```

```r
stan_dev2
```

```
## [1] 0.03183784
```

```r
## Comparing the two exercises
comparison <- data.frame(
  Sample_size = c(20, 1000),
  Average_cor = c(avg_cor, avg_cor2),
  Standard_Dev = c(stan_dev, stan_dev2)
)

comparison
```

```
##   Sample_size  Average_cor Standard_Dev
## 1          20 -0.002123023   0.23073518
## 2        1000  0.003448129   0.03183784
```

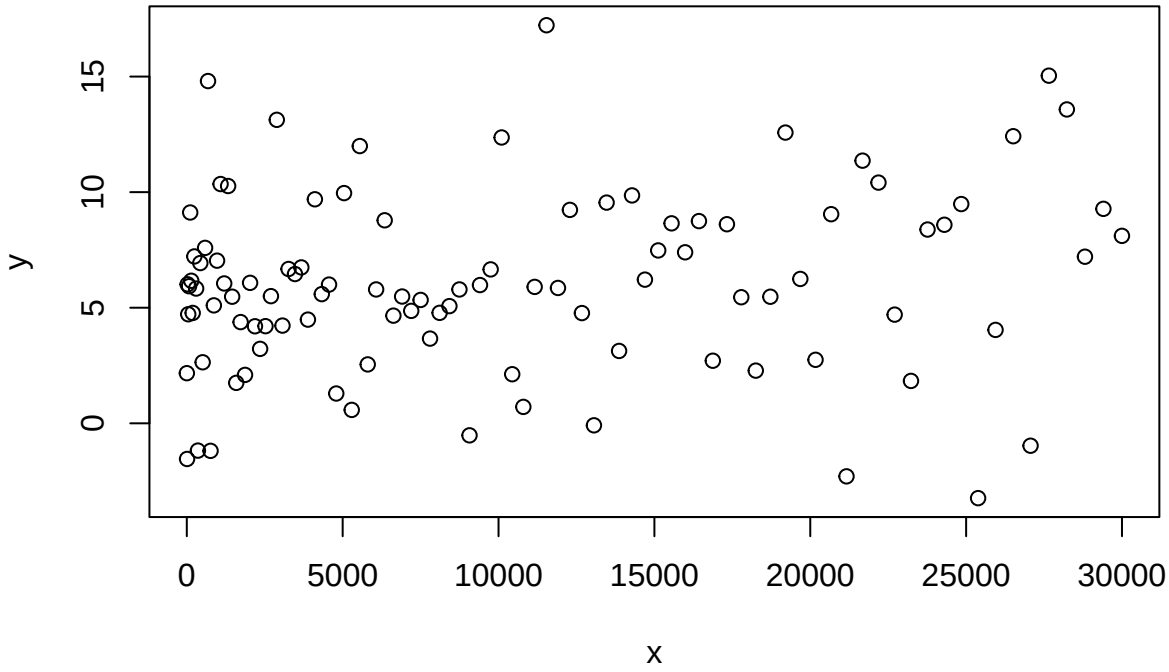
As can be seen in the table above, the sample estimate for the correlation is much closer to the population correlation, which should be 0. Additionally, the standard deviation has also shrunk considerably, which means there is less variance in the sampling distribution of the correlation. Therefore, we can expect our sample estimate to be much more accurate with a larger sample size.

Question 3

```
res2 <- data.frame(x = numeric(), y = numeric()) #Creating an empty dataset


z <- rnorm(100,mean=0,sd=3) #z contains 100 values drawn from a normal distribution
x <- z + 3*(sample(1:100))^2 #Creating variable x as z plus random noise
y <- z + (sample(1:100))^(1/2) #Doing the same with y

plot(x, y) #Creating a scatter plot to visualize the relationship between x and y.
```




```
cor_val3 <- cor(x, y) #Getting the correlation between x and y
cor_val3
```

```
## [1] 0.1669864
```

This shows that we can find a correlation between two variables when neither is directly linked to each other but are instead linked to a third confounding variable. It implies that just because two variables exhibit correlation, it does not necessarily imply causation.

A famous example of such a spurious correlation is the apparent correlation between the amount of ice creams consumed and the number of shark attacks throughout a year. However, this can be explained by people being more likely to consume ice creams and go for swims in the beach during the summer months. In this case, summer is the counfounding factor that can give the impression of there being a spurious correlation between the the amount of ice creams consumed and the number of shark attacks throughout a year.