Problem Set 2

Fernanda Valdez

2025-10-22

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
require(ggplot2)
set.seed(5315)
```

Simulation

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?

Two random variables with 20 observations each:

```
a <- rnorm(20)
b <- rnorm(20)
```

Calculate correlation between the two variables:

```
cor(a, b)
```

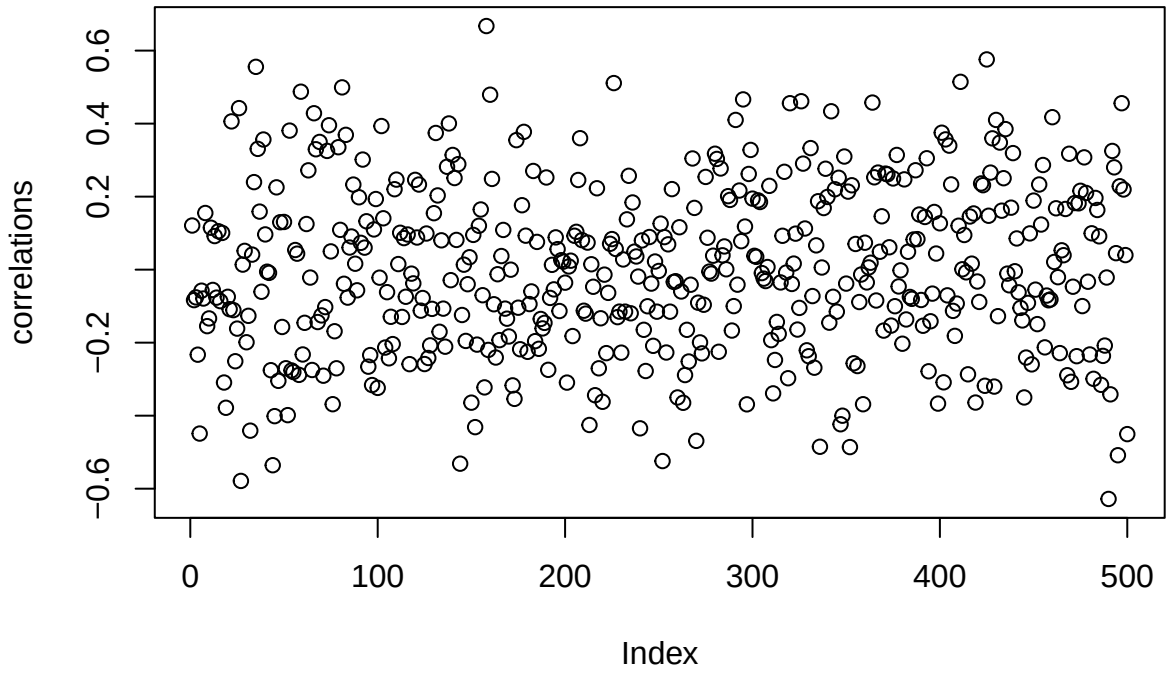
```
## [1] -0.1871898
```

Repeat this process many times:

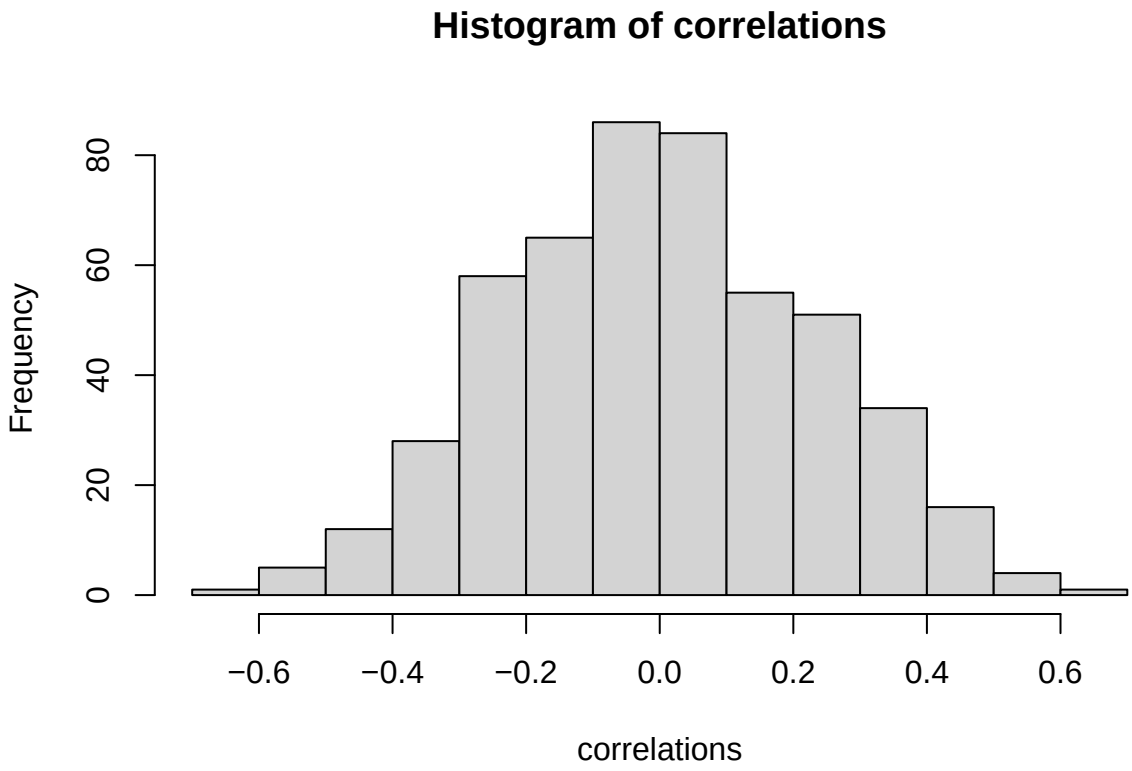
```
correlations <- replicate(500,{
  a <- rnorm(20)
  b <- rnorm(20)
  cor(a,b)})
```

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

```r
plot(correlations)
```




```r
hist(correlations)
```




```r
sd(correlations)
```

```
## [1] 0.2290531
```

On average, we would expect these two variables to be close to 0 because they are two independent random

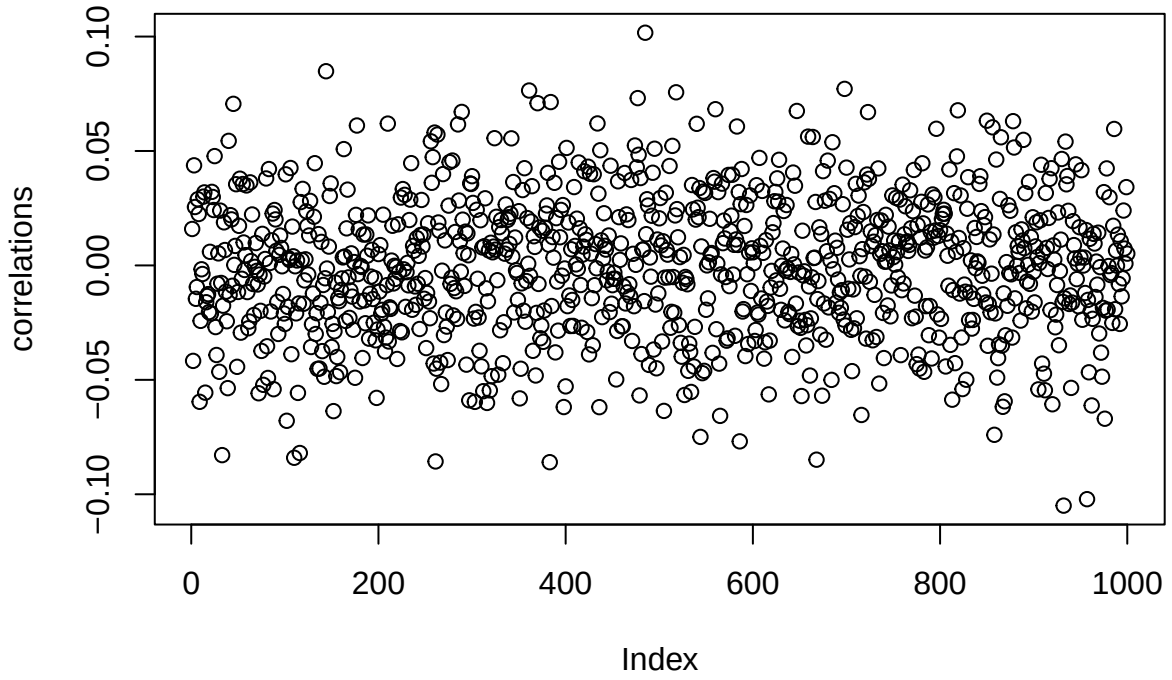
variables. This small distribution can mistakenly allude you to think there is a relationship.

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

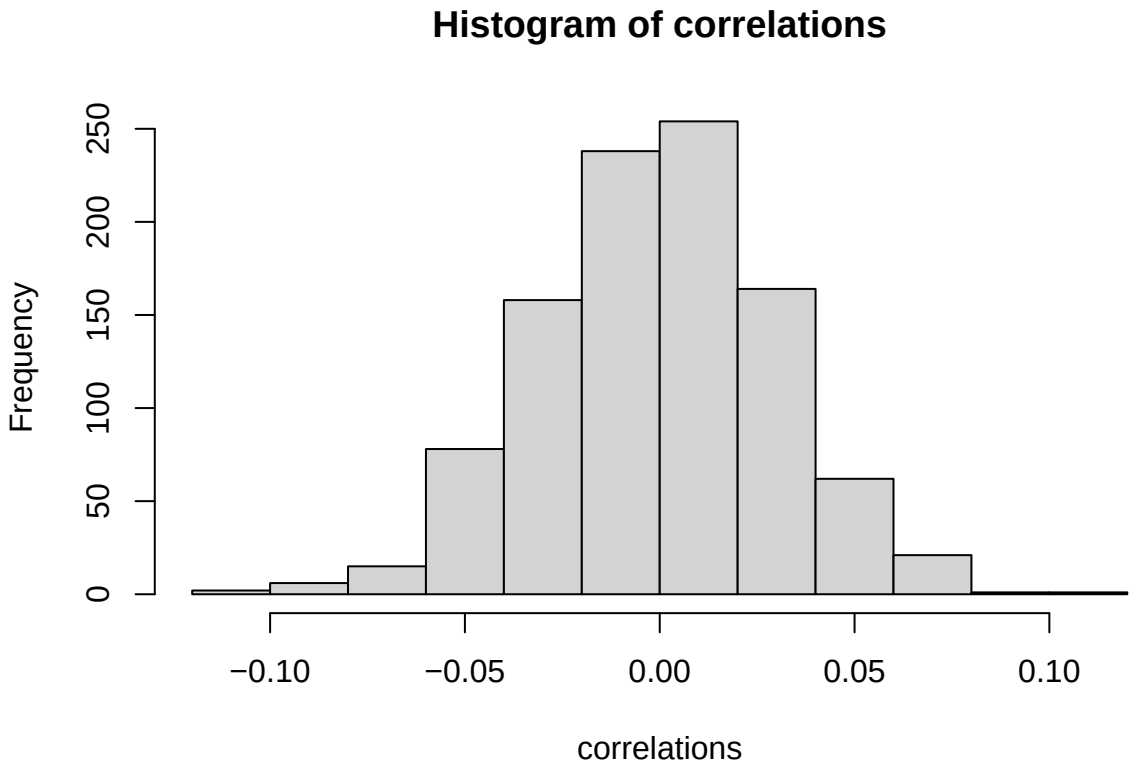
```
correlations <- replicate(1000,{
  a <- rnorm(1000)
  b <- rnorm(1000)
  cor(a,b)})

sd(correlations)
```

```
## [1] 0.03014421
```

```
plot(correlations)
```




```
hist(correlations)
```



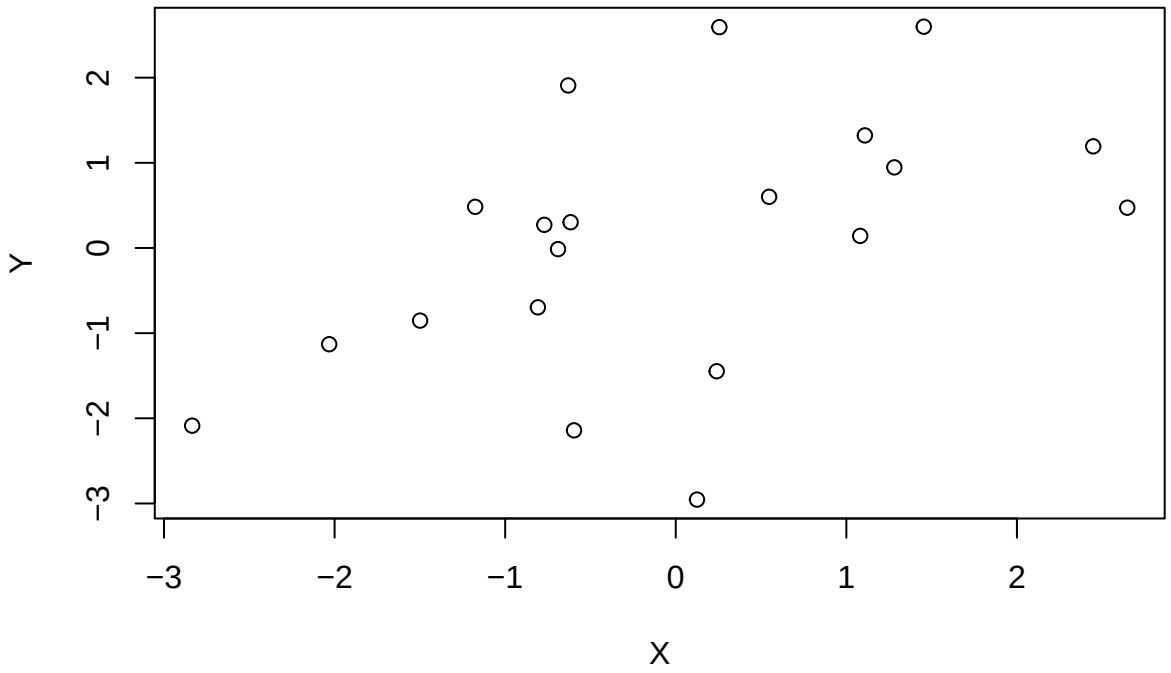
Here, the results differ in that because the sample size is much larger, it is able to better illustrate that there is no correlation between the two variables. A fact that a smaller sample as previously shown will hide.

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

```
Z <- rnorm(20)
X <- Z + rnorm(20)
Y <- Z + rnorm(20)

cor (X,Y)
```

```
## [1] 0.4922961
```

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



This tells us that because X and Y both have a relationship with Y, they are still going to have some correlation between them given their respective relationships with Y.