Problem Set 2

Ellie Choe

2025-10-11

—- Simulation —-

Creat two random variables (n = 20) and compute their correlation

```r
set.seed(123)
n <- 20
x <- rnorm(n)
y <- rnorm(n)
cor(x, y)
```

```
## [1] -0.09172278
```

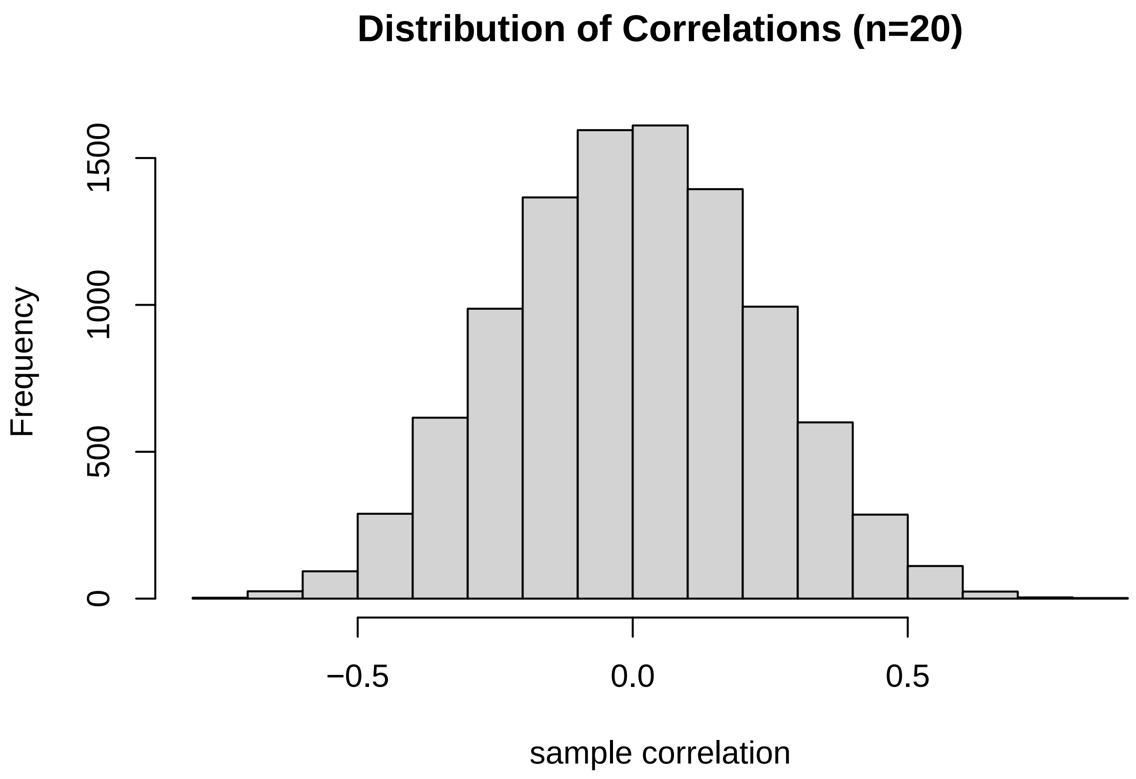
Repeat this process many times

```r
n_sim <- 10000

cors <- replicate(n_sim, {
  x <- rnorm(n)
  y <- rnorm(n)
  cor(x, y)
})
```

Plot the distribution and report summary stats

```r
hist(cors, main = "Distribution of Correlations (n=20)", xlab = "sample correlation")
```



```
sd(cors)
```

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

```
mean(cors)
```

```
## [1] 0.001348848
```

Interpretation:

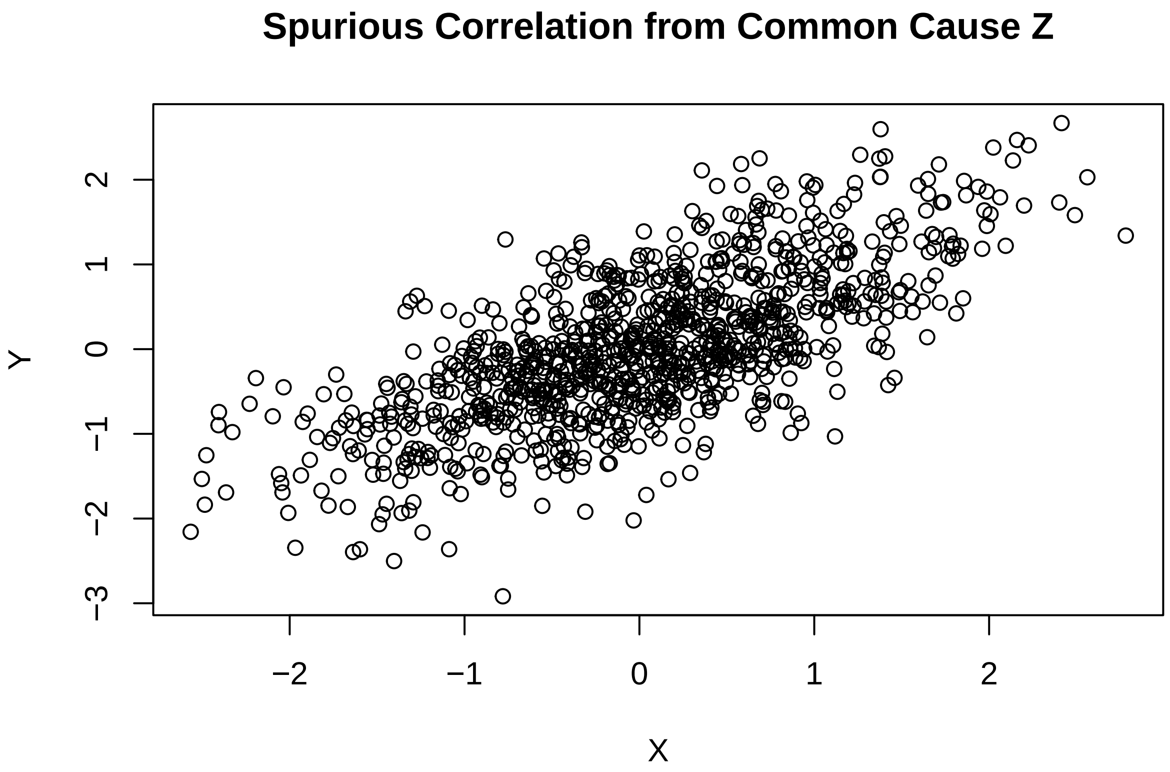
It shows that when the sample size is small, the sample correlation can vary widely even if the true population correlation is 0. In other words, sample estimates are noisy and unstable for small samples because of random fluctuations due to limited data.

Repeat with a larger sample (n = 1000)

```
n <- 1000
Z <- rnorm(n)
X <- 0.7 * Z + rnorm(n, sd = 0.5)
Y <- 0.7 * Z + rnorm(n, sd = 0.5)
```

Plot X and Y to show spurious correlation

```
plot(X, Y, main = "Spurious Correlation from Common Cause Z")
```



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

```
## [1] 0.7025198
```