prob set 2

2025-10-16

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
```

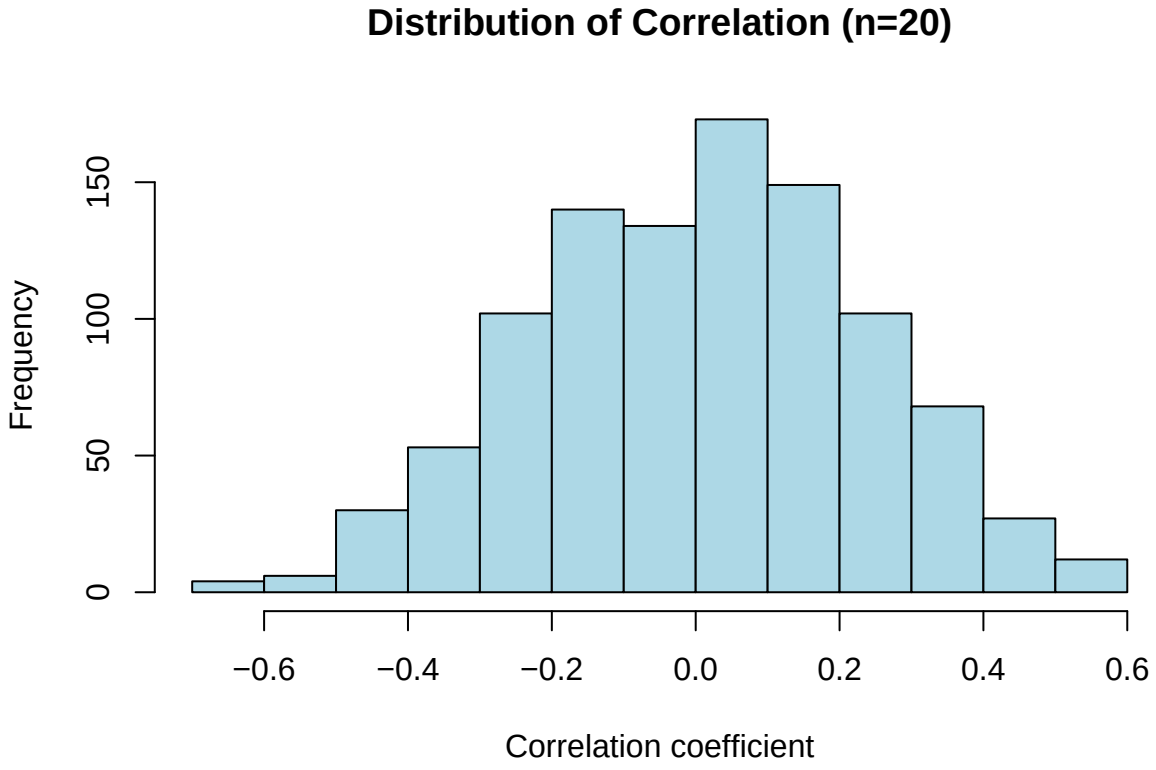
1.

```
#set seed for reproductiability
set.seed(121)

#set numbers of observations and iterations
n <- 20
n_iter <- 1000

#simulate the correlation coefficient and save in a vector
correlation <- replicate(n_iter,{
  x <- rnorm(n)
  y <- rnorm(n)
  cor(x,y)
})

#draw the histogram of correlation coefficient
hist_c <- hist(correlation,
     col = "lightblue",
     main = "Distribution of Correlation (n=20)",
     xlab = "Correlation coefficient")
```




```
mean_corr <- mean(correlation)
print(mean_corr)
```

```
## [1] 0.007809592
```

```
sd_corr <- sd(correlation)
print(sd_corr)
```

```
## [1] 0.2313069
```

On average we would expect the correlation coefficient between the two variable to be zero. This is because they are created independently by rnorm(), so the two variables are independent, although they share the same mean and standard deviation. For two independent variables, their covariance is zero, so the correlation coefficient is also zero.

However, the standard deviation of the correlation coefficients is approximately 0.2313. It means for a sample size of n=20, the correlation coefficients tend to deviate from the population parameters by around 0.2313 on average.

It tells us the existence of the sampling error, a random noise within the sample estimate when the sample is not large enough.

2.

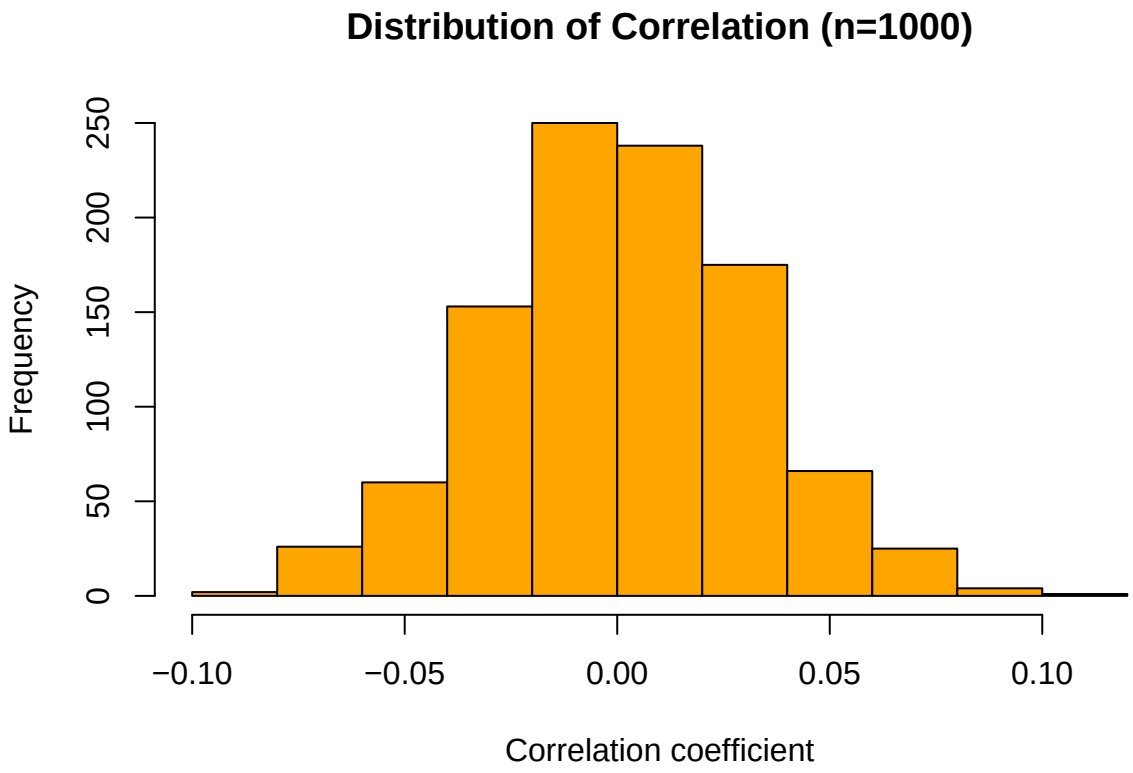
```
#set seed for reproductiability
set.seed(121)

#set new numbers of 1000 observations
n1 <- 1000
n_iter <- 1000

#simulate the correlation coefficient and save in a vector
```

```r
correlation1 <- replicate(n_iter,{
  x1 <- rnorm(n1)
  y1 <- rnorm(n1)
  cor(x1,y1)
})

#draw the histogram of correlation coeffcient
hist_c1 <- hist(correlation1,
     col = "orange",
     main = "Distribution of Correlation (n=1000)",
     xlab = "Correlation coefficient")
```




```r
mean_corr1 <- mean(correlation1)
print(mean_corr1)
```

```
## [1] 0.001041378
```

```r
sd_corr1 <- sd(correlation1)
print(sd_corr1)
```

```
## [1] 0.03061228
```

After increasing the sample size to 1000 observations, there is a substantial difference between the results.

Firstly, the average correlation coefficient decreases from approximately 0.0078 to 0.001, which means the average correlation approaches to our expectation of 0 between two independent variables.

Secondly, the standard deviation of correlation coefficients decreases significantly from approximately 0.2313 to 0.0306, which means the correlation coefficients tend to deviate less from the population parameter. We can also see this from the histogram, as the spread of correlation coefficient on the x-axis narrowed from a range of +-0.6 to +-0.1.

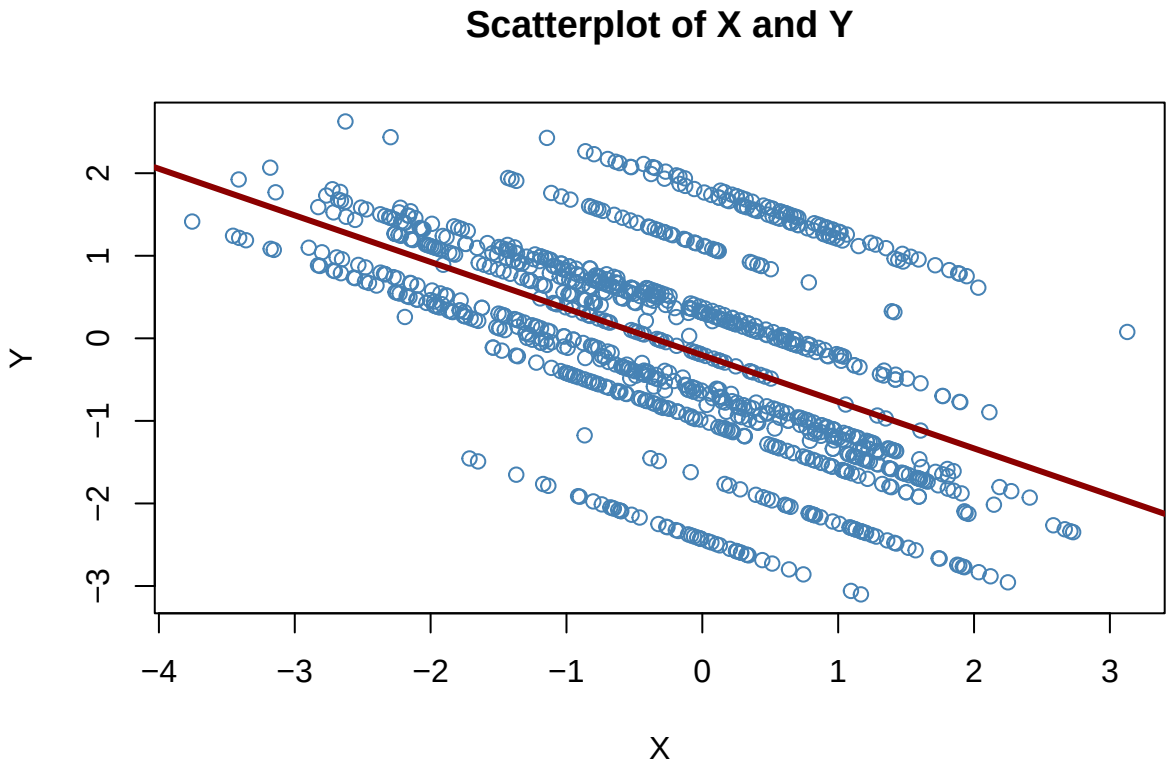
3.

```r
#set seed for reproductiability
set.seed(121)

#create RV Z with 1000 obs
 Z <- rnorm(n1)

#create RV X and Y as functions of Z plus random noise
 X <- 0.7*Z + rnorm(n)
 Y <- -0.4*Z + rnorm(n)

#make scatterplot
 plot(X, Y, main = "Scatterplot of X and Y",
      xlab = "X", ylab = "Y",
      col = "steelblue")
 abline(lm(Y ~ X), col = "darkred", lwd = 3)
```




```r
#report correlation coefficient
 cor(X,Y)
```

```
## [1] -0.5492345
```

This tells us we need to be cautious about spurious correlation when interpreting correlations. Specifically, although the scatter plot and the correlation coefficient show that there is a strong negative correlation between X and Y (with coefficient -0.549), but X and Y actually have no causal relationship. It is the confounder Z which causes both X and Y leads to this spurious correlation. It also tells us that correlation does not imply causation.