Problem_Set_2

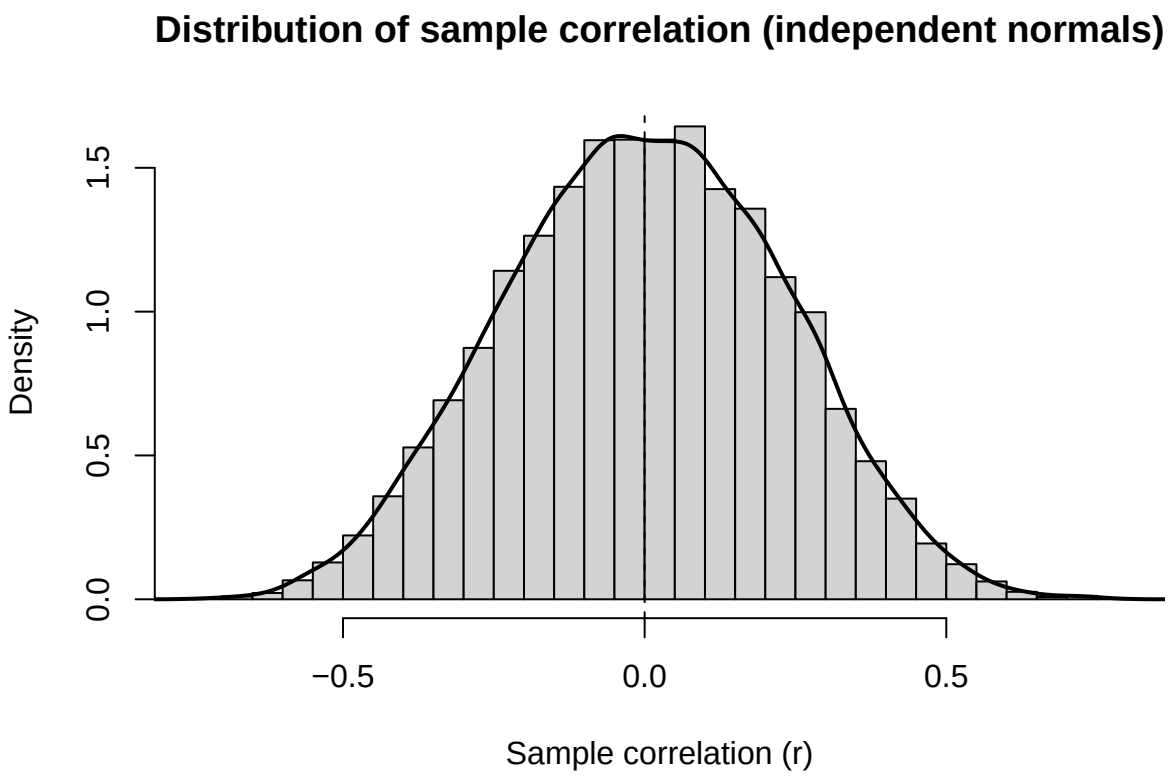
Santiago Vidal

2025-10-23

```r
set.seed(1234)

# 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
# > coeIicients 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?

# parameters
n <- 20
B <- 10000 # number of Monte Carlo replications

# function that returns the sample correlation from two independent N(0,1)
sim_once <- function(n) {
  x <- rnorm(n)
  y <- rnorm(n)
  cor(x, y)
}

# run the simulation
cors_n20 <- replicate(B, sim_once(n))

# summary statistics
mean_n20 <- mean(cors_n20)
sd_n20 <- sd(cors_n20)
quantile_n20 <- quantile(cors_n20, probs = c(.01, .05, .10, .25, .5, .75, .90, .95, .99))

list(mean = mean_n20, sd = sd_n20, quantiles = quantile_n20)
```

```
## $mean
## [1] 0.0005750106
##
## $sd
## [1] 0.2294137
##
## $quantiles
##           1%           5%          10%          25%          50%          75%
## -0.511594186 -0.379600574 -0.301876818 -0.160915854  0.002096046  0.164247685
##          90%          95%          99%
##  0.296022606  0.373363941  0.509617885
```

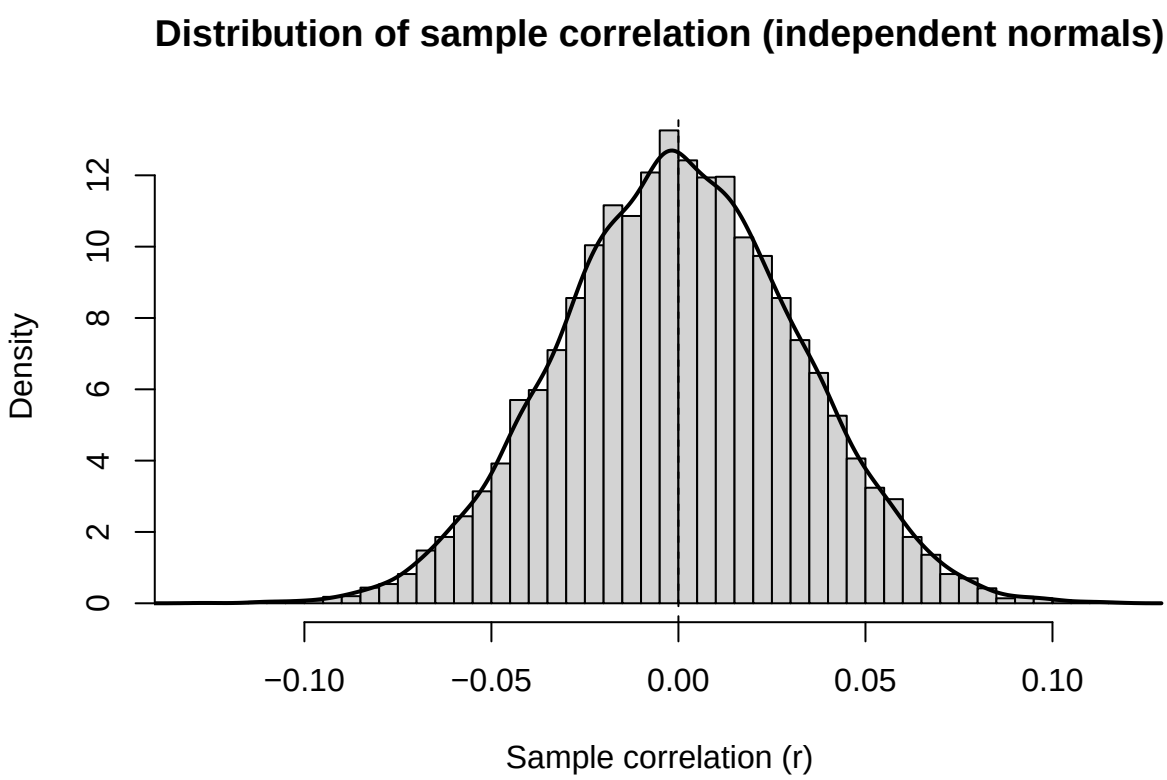
```r
hist(cors_n20, breaks = 40, probability = TRUE,
     main = "Distribution of sample correlation (independent normals)",
     xlab = "Sample correlation (r)")
lines(density(cors_n20), lwd = 2)
abline(v = 0, lty = 2)
```




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

n_large <- 1000
cors_n1000 <- replicate(B, sim_once(n_large))

mean_n1000 <- mean(cors_n1000)
sd_n1000 <- sd(cors_n1000)
quantile_n1000 <- quantile(cors_n1000, probs = c(.01, .05, .10, .25, .5, .75, .90, .95, .99))

list(mean = mean_n1000, sd = sd_n1000, quantiles = quantile_n1000)
```

```
## $mean
## [1] 0.0002157789
##
## $sd
## [1] 0.03173315
##
## $quantiles
##            1%            5%           10%           25%           50%
## -7.238442e-02 -5.203578e-02 -4.069465e-02 -2.127657e-02  2.609834e-06
##           75%           90%           95%           99%
##  2.163611e-02  4.118685e-02  5.261325e-02  7.285291e-02
```

```r
hist(cors_n1000, breaks = 40, probability = TRUE,
     main = "Distribution of sample correlation (independent normals)",
     xlab = "Sample correlation (r)")
lines(density(cors_n1000), lwd = 2)
abline(v = 0, lty = 2)
```

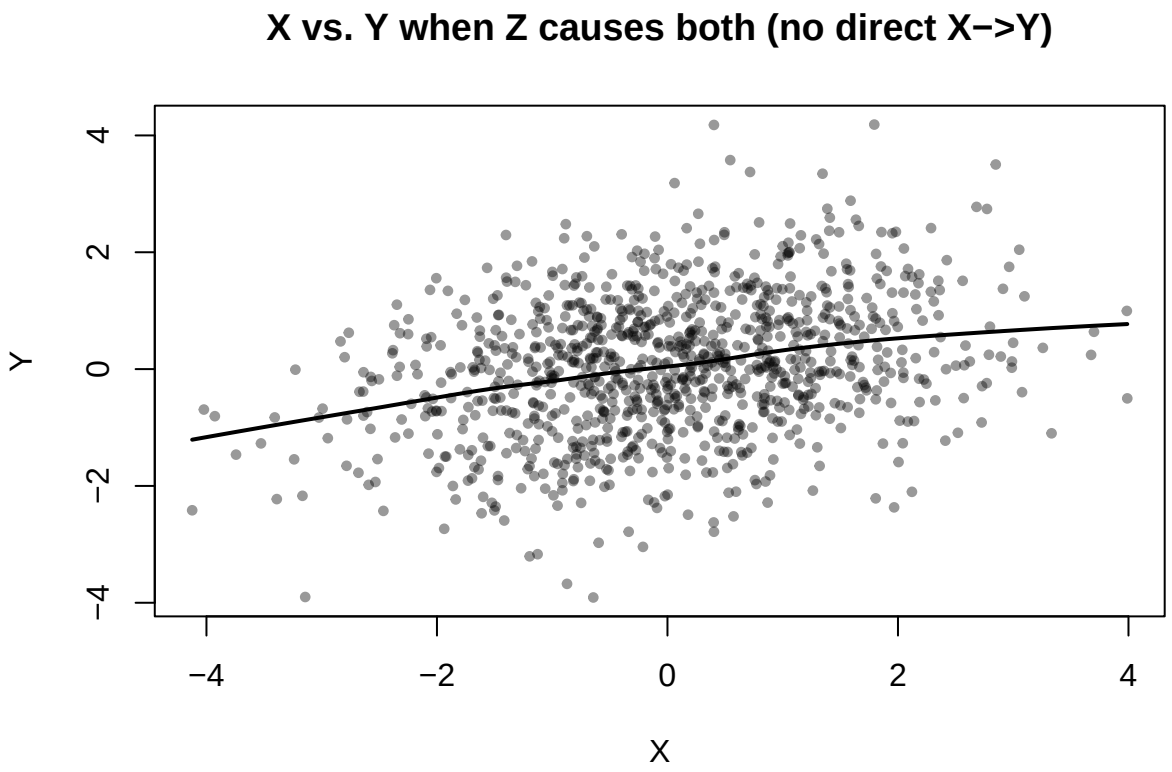



```r
# 3. Create three random variables in R that have the following causal
# > relationship: That is, Z causes both X and Y, but X and Y have no causal
# > relationship. Plot X and Y on a scatter plot and report their correlation.
# > What does this tell us about interpreting correlations?
# > Hint: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

N <- 1000

# Generate Z and then X, Y as functions of Z with independent noise
Z <- rnorm(N)

# Strength of Z's effect on X and Y
alpha <- 0.8
beta <- 0.6

# Independent noise terms
sx <- 1
sy <- 1

X <- alpha * Z + rnorm(N, sd = sx)
Y <- beta * Z + rnorm(N, sd = sy)
```

```r
# Compute the correlation between X and Y
cor_xy <- cor(X, Y)
cor_xy
```

```
## [1] 0.2960032
```

```r
plot(X, Y, pch = 19, cex = 0.6, col = rgb(0, 0, 0, 0.4),
     xlab = "X", ylab = "Y", main = "X vs. Y when Z causes both (no direct X→Y)")
# simple smooth for visualization
lines(lowess(X, Y, f = 2/3, iter = 1), lwd = 2)
```




```r
# Interpretation: Even though there is no direct causal path from X to Y,
# > they appear correlated because they share a common cause (Z).
# > This is an example of a spurious correlation due to confounding effects.
```