POLS 6O2-PS2-Joseph Kim

Joseph Kim

2025-10-22

Q1: correlation distribution with n=20

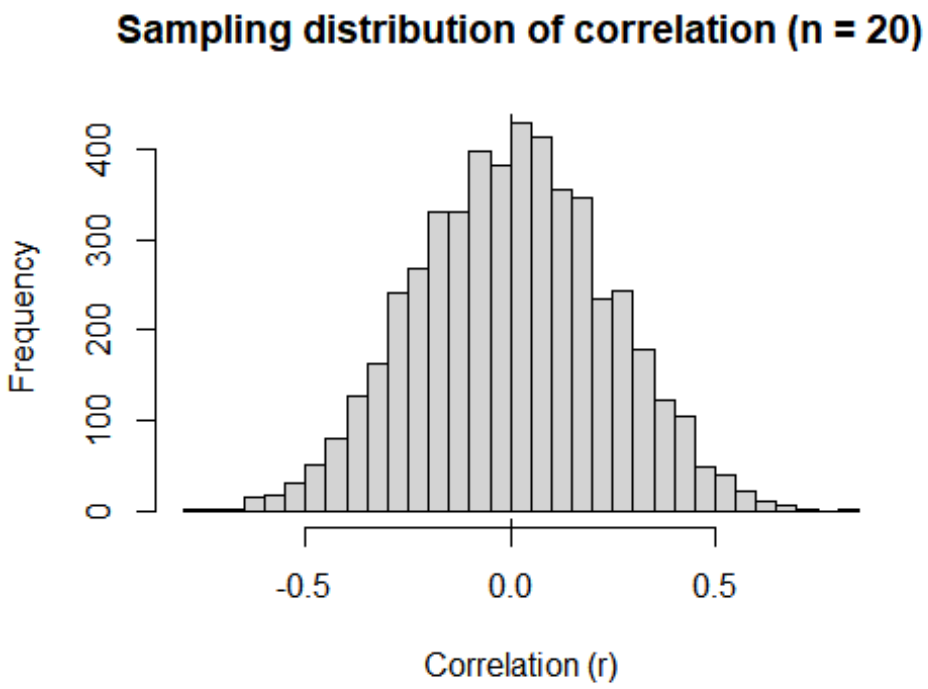
```r
set.seed(123)

n <- 20               # sample size
M <- 5000             # number of repetitions (increase if you want)
cors <- numeric(M)    # container for correlations

for (m in 1:M) {
  x <- rnorm(n)       # X ~ N(0,1)
  y <- rnorm(n)       # Y ~ N(0,1), independent of X
  cors[m] <- cor(x, y)
}

# Summary stats
q1_mean <- mean(cors)                    # expected ~ 0
q1_sd   <- sd(cors)                      # sampling SD of r at n=20

cat("Q1 mean(corr) =", round(q1_mean, 4), "\n")
```

```
## Q1 mean(corr) = 0.0044
```

```r
cat("Q1 sd(corr)   =", round(q1_sd,   4), "\n")
```

```
## Q1 sd(corr)   = 0.2348
```

```r
# Histogram of the sampling distribution of r
hist(cors, breaks = 40,
     main = "Sampling distribution of correlation (n = 20)",
     xlab = "Correlation (r)")
abline(v = 0, lty = 2)
```



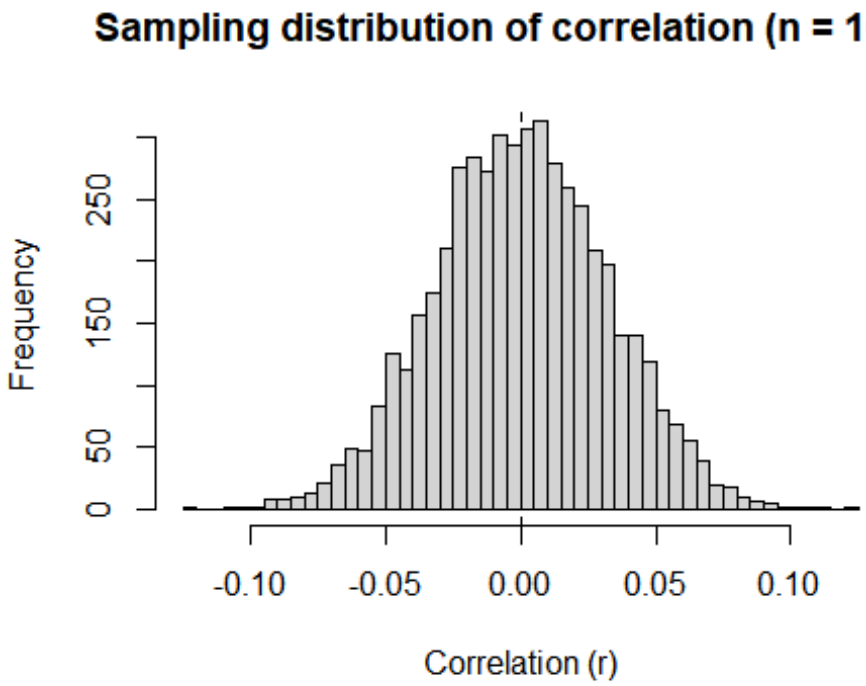
Interpretation On average, the correlation is approximately 0 (0.0044), because X and Y are independent. With n = 20, the distribution is wide—sample correlations vary substantially due to random sampling.(sd=0.2348) This demonstrates that small samples produce noisy estimates of population parameters.

Q2: correlation distribution with n=1000

```r
set.seed(123)

n2 <- 1000
M2 <- 5000
cors2 <- numeric(M2)

for (m in 1:M2) {
  x <- rnorm(n2)
  y <- rnorm(n2)
  cors2[m] <- cor(x, y)
}

q2_mean <- mean(cors2)
q2_sd   <- sd(cors2)

cat("Q2 mean(corr) =", round(q2_mean, 6), "\n")
```

```
## Q2 mean(corr) = 0.000361
```

```r
cat("Q2 sd(corr)   =", round(q2_sd,   6), "\n")

## Q2 sd(corr)   = 0.032084

hist(cors2, breaks = 40,
     main = "Sampling distribution of correlation (n = 1000)",
     xlab = "Correlation (r)")
abline(v = 0, lty = 2)
```

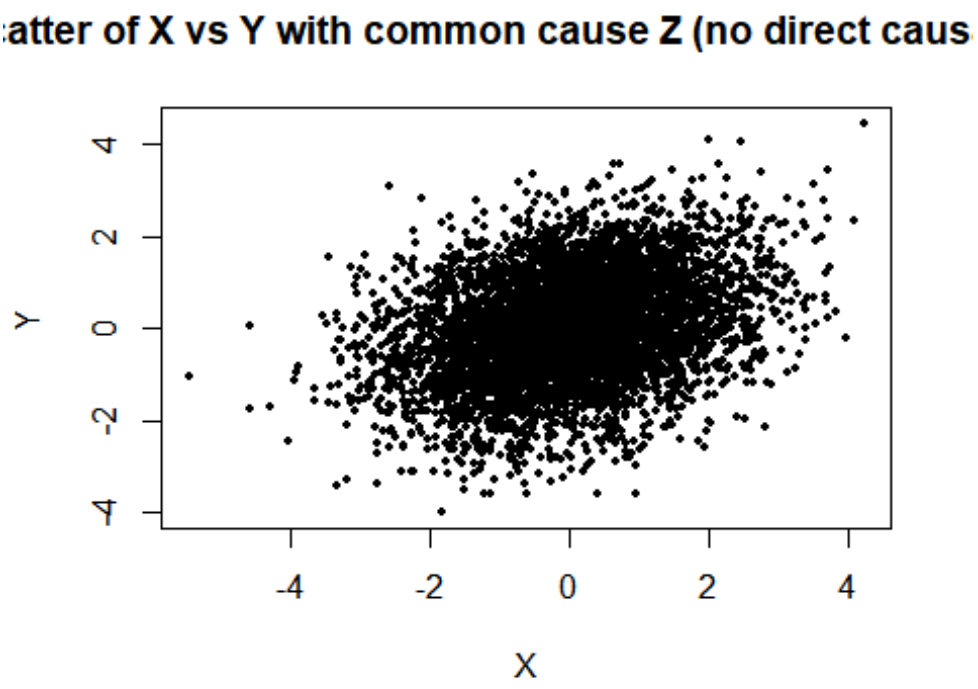


Interpretation The mean starys nearly the same around 0 (0.000361). That said, the standard deviation is much smaller (0.2348 vs 0.032084), and is more tightly concentrated near 0. This indicates as the sample size gets larger, sampling variability diminishes. In other words, the precision of estimation grows with larger sample size.

Q3: Confounding via common cause Z

```r
set.seed(123)

N <- 5000

# Common cause
Z <- rnorm(N, mean = 0, sd = 1)

# X and Y each depend on Z + their own noise; no direct X->Y or Y->X effect
a <- 0.8  # strength of Z -> X
b <- 0.6  # strength of Z -> Y
```

```r
X <- a*Z + rnorm(N, 0, 1)   # X = a*Z + εx
Y <- b*Z + rnorm(N, 0, 1)   # Y = b*Z + εy

# Correlation and scatter
r_xy <- cor(X, Y)
cat("Q3 cor(X, Y) =", round(r_xy, 4), "\n")

## Q3 cor(X, Y) = 0.3207

plot(X, Y, pch = 16, cex = 0.6,
     main = "Scatter of X vs Y with common cause Z (no direct causal link)",
     xlab = "X", ylab = "Y")
```



Interpretation The correlation between X and Y does seem to exist, albeit weak and fuzzy. However, there is no causal relationship between the two. This ostensible correlation is spurious, and stems from the fact that Z influences both X and Y. The bottom line is that correlation is not causation. A confounder can induce correlation without a direct effect between X and Y.