Hamed Nikookalam

Table of Contents

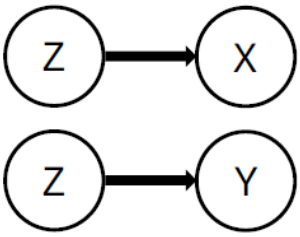
Problem Set 2

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?

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

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.

Solution

1. On average, the correlation between the two variables is 0 because we generated them to be unrelated; the histogram of sample correlations shows how much the sample estimate can bounce around the true value just from random sampling, especially with small samples. For n = 20, this sampling variability has a standard deviation of about 0.23, so you can easily see moderately positive or negative correlations purely by chance; as n grows, the distribution tightens around 0 and the estimate becomes more precise.

2. Both cases (n:20 and n:1000) have an expected correlation of 0 because the variables are unrelated, but the larger sample produces far more precise estimates—narrower confidence intervals and a much lower risk of mistaking random noise for a real relationship.

3. Z drives both X and Y, but there is no direct causal path between X and Y; nevertheless, correlation is typically positive because X and Y both inherit variation from the same Z. The scatter plot shows an apparent linear trend even though X does not cause Y (and vice versa). This demonstrates that a common cause can create correlation between two variables with no direct causal relationship, so correlation by itself should not be interpreted as causation.

Code

```
# Q1
set.seed(1)
# 20 observations for each variable
n_x <- 20
n_y <- 20
# Repeat many times
B   <- 10000
# sample correlation for this run
rvals <- replicate(B, {
  x <- rnorm(n_x)
  y <- rnorm(n_y)
  cor(x, y)
})
# SD
sd(rvals)
hist(rvals, breaks = 40, xlab = "r",
     main = "n_x = n_y = 20 (Sampling distribution)")
# drawing a vertical line at 0
abline(v = 0, lwd = 4)
##.....
# Q2
set.seed(1)
# 1000 observations for each variable
n_x <- 1000
n_y <- 1000
# Repeat many times
B   <- 10000
```

```
# sample correlation
rvals_big <- replicate(B, {
  x <- rnorm(n_x)
  y <- rnorm(n_y)
  cor(x, y)
})
# SD
sd(rvals_big)
hist(rvals_big, breaks = 40, xlab = "r",
     main = "n_x = n_y = 1000 (Sampling distribution)")
abline(v = 0, lwd = 4)
#.....................................................
# Q3
# Sample size
n <- 500
# Common cause (Z)
Z <- rnorm(n)
# X and Y are functions of Z
X <- Z + rnorm(n)
Y <- Z + rnorm(n)
# Scatter plot of X vs Y
plot(X, Y, pch = 19, cex = 0.6, xlab = "X", ylab = "Y")
# Correlation between X and Y
cor(X, Y)
```