A Simple Way to Generate a Correlated Binary Variable in R

Danni Shi

December 2022

Given a binary variable $Z_1 \sim \text{Bernoulli}(p_1)$, how can we can generate a binary variable $Z_2 \sim \text{Bernoulli}(p_2)$ such that $\text{Corr}(Z_1, Z_2) = \rho$? First, we have

$$\begin{aligned}
\text{Corr}(Z_1, Z_2) &= \frac{\text{Cov}(Z_1, Z_2)}{\sqrt{\text{Var}(Z_1)\text{Var}(Z_2)}} \\
&= \frac{E(Z_1 Z_2) - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} = \rho \\
\Rightarrow E(Z_1 Z_2) &= \rho\sqrt{p_1(1-p_1)p_2(1-p_2)} + p_1 p_2 \\
&= \text{Pr}(Z_1 = 1, Z_2 = 1)
\end{aligned}$$

Let $E(Z_1 Z_2) = \text{Pr}(Z_1 = 1, Z_2 = 1) = s$, note that

$$\begin{aligned}
\text{Pr}(Z_2 = 1 | Z_1 = 1) = E(Z_2 | Z_1 = 1) = \frac{\text{Pr}(Z_2 = 1, Z_1 = 1)}{\text{Pr}(Z_1 = 1)} = \frac{s}{p_1} \\
\text{Pr}(Z_2 = 1 | Z_1 = 0) = E(Z_2 | Z_1 = 0) = \frac{\text{Pr}(Z_2 = 1, Z_1 = 0)}{\text{Pr}(Z_1 = 0)} = \frac{p_2 - s}{1 - p_1}
\end{aligned}$$

So we have

$$Z_2 | Z_1 = 1 \sim \text{Bernoulli}\left(\frac{s}{p_1}\right); \; Z_2 | Z_1 = 0 \sim \text{Bernoulli}\left(\frac{p_2 - s}{1 - p_1}\right)$$

Here is an example and a function of how we can generate a binary variable that is correlated to an existing binary variable.

```r
n <- 1000
p1 <- 0.5
Z1 <- rbinom(n,1,p1) # given and fixed

correlated.binary <- function(n,p1,p2,rho,Z1) {
  Z2 <- rep(NA,n)
  s <- rho*sqrt(p1*(1-p1)*p2*(1-p2))+p1*p2

  Z1.ones <- which(Z1==1)
  Z2[Z1.ones] <- rbinom(sum(Z1==1),1,s/p1)

  Z1.zeros <- which(Z1==0)
  Z2[Z1.zeros] <- rbinom(sum(Z1==0),1,(p2-s)/(1-p1))
  return(Z2)
}

p2 <- 0.3
```

```
rho <- 1/3
Z2 <- correlated.binary(n,p1,p2,rho,Z1)
mean(Z2)
```

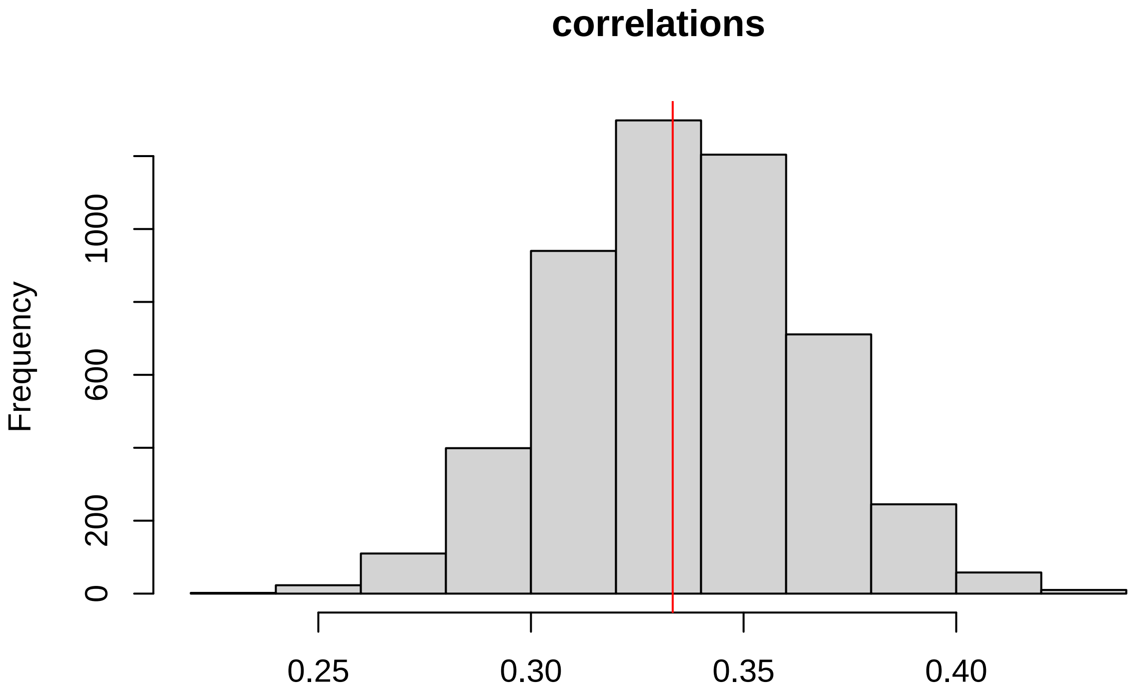
```
## [1] 0.301
```

```
cor(Z1,Z2)
```

```
## [1] 0.3897139
```

```
N <- 5000
cors <- rep(NA,N)
for (i in 1:N) {
  Z2 <- correlated.binary(n,p1,p2,rho,Z1)
  cors[i] <- cor(Z1,Z2)
}

hist(cors,main="correlations",xlab=""); abline(v=rho, col="red")
```


correlations
Frequency