Correlation and Regression
Suppose we have paired data \((X_1, Y_1), \ldots, (X_n, Y_n)\).

Pearson’s correlation coefficient coefficient is

\[
\hat{r}_{XY} \equiv \frac{\hat{\text{cov}}(Y, X)}{SD(Y)SD(X)} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum (Y_i - \bar{Y})^2 \cdot \sum (X_i - \bar{X})^2}}.
\]

It is a fact that \(-1 \leq r_{XY} \leq 1\), and \(r_{XY} = 1\) only if the \((X_i, Y_i)\) pairs lie exactly on a line.

If \(r_{XY} = 0\) then \(X\) and \(Y\) are uncorrelated. This means there is no systematic linear relationship between \(X\) and \(Y\). It does not imply that \(X\) and \(Y\) are independent.
To simulate data with a given population correlation coefficient $r$, first generate two independent random vectors $X$ and $U$ with population mean zero and population variance one. Then set

$$Y = rX + \sqrt{1 - r^2}U.$$ 

We can calculate that

$$\text{var}(Y) = r^2\text{var}(X) + (1 - r^2)\text{var}(U) + 2r\sqrt{1 - r^2}\text{cov}(X,U) = 1$$

and

$$\text{cov}(X,Y) = \text{cov}(X, rX + \sqrt{1 - r^2}U) = r\text{var}(X) + \sqrt{1 - r^2}\text{cov}(X,U) = r.$$
The following program assesses the bias, variance and MSE of the sample correlation coefficient for the population correlation coefficient.

```r
## Population correlation coefficients to consider.
R = seq(-0.9, 0.9, 0.1)

## Sample size.
n = 20

## Number of simulation replications.
nrep = 1e3

B = array(0, c(length(R),4))

## Loop over the different population correlation coefficients.
for (j in 1:length(R)) {
  r = R[j]
  
  ## Generate nrep sample correlation coefficients.
  for (i in 1:nrep) {
    ...
  }
}
```
C = array(0, nrep)
for (i in 1:nrep) {
  X = rnorm(n)
  Y = r*X + sqrt(1-r^2)*rnorm(n)
  C[i] = cor(X,Y)
}

## Bias, MSE, and variance.
B[j,] = c(r, mean(C)-r, mean((C-r)^2), var(C))
Suppose we want to produce a confidence interval for the population correlation coefficient, based on the sample correlation coefficient.

Fisher’s transform is the most standard approach. The following statistic approximately follows a normal distribution, regardless of the population correlation coefficient.

\[ Z = \frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right). \]

The expected value of \( Z \) is approximately

\[ \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right), \]

and the variance is approximately \( 1/(n - 3) \).
To form a confidence interval, start with

\[
P\left(-2/\sqrt{n-3} \leq \frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right) - \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) \leq 2/\sqrt{n-3} \right) \approx 0.95.
\]

Therefore the CI is

\[
\frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right) - 2/\sqrt{n-3} \leq \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) \leq \frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right) + 2/\sqrt{n-3}.
\]

Letting

\[
C_L = \frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right) - 2/\sqrt{n-3} \quad \quad C_U = \frac{1}{2} \log \left( \frac{1 + \hat{r}}{1 - \hat{r}} \right) + 2/\sqrt{n-3}
\]
We get

\[ P(C_L \leq \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) \leq C_U) \approx 0.95. \]

The inverse of Fisher’s transform is

\[ \frac{\exp(2Z) - 1}{\exp(2Z) + 1}, \]

therefore

\[ P \left( \frac{\exp(2C_L) - 1}{\exp(2C_L) + 1} \leq r \leq \frac{\exp(2C_U) - 1}{\exp(2C_U) + 1} \right) \approx 0.95. \]