

Paired Data

Let our data be $(X_1, Y_1), \dots, (X_n, Y_n)$, where:

$$Y_i = rX_i + \sqrt{1-r^2}U_i$$

Where the X_i 's and U_i 's are iid standard normal. Let's verify that the approximate 95 % CI for r given on page 7 of Dr. Shedden's lecture notes has roughly 95 % coverage probability. Consider a sequence of population correlation coefficients, and a sample size of $n = 20$.

```
## Population correlation coefficients to consider.
```

```
R = seq(-0.9, 0.9, 0.1)
```

```
## Sample size.
```

```
n = 20
```

```
## Number of simulation replications.
```

```
nrep = 1e3
```

```
PROP=NULL
```

```
## Loop over the different population correlation coefficients.
```

```
for (j in 1:length(R))
```

```
{
```

```
  r = R[j]
```

```
  ## Generate nrep sample correlation coefficients.
```

```
  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)
```

```
  }
```

```
  ## Generate nrep approximate 95% CI's
```

```
  logterm = 0.5*log(1+C)-0.5*log(1-C)
```

```
  CL = logterm-2/sqrt(n-3)
```

```
  CU = logterm+2/sqrt(n-3)
```

```
  LB = (exp(2*CL)-1)/(exp(2*CL)+1)
```

```
  UB = (exp(2*CU)-1)/(exp(2*CU)+1)
```

```
  ## Find the proportion of population pearson-correlation coeffs
```

```

## that are contained in the 95 percent CI for this value value
## of r

PROP = c(PROP, mean( (LB <= r) & (UB >= r) ))
}

## Look at the output
print(cbind(R, PROP))

```

Now let's use this confidence interval to compute the power of the test of the hypothesis that the population Pearson correlation coefficient r is nonzero. Consider the same sequence of population Pearson correlation coefficients used above. Note that all we need to do is to find the proportion of CI's not including zero for each value of r .

```

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

## Sample size.
n = 20

## Number of simulation replications.
nrep = 1e3

POWER=NULL

## Loop over the different population correlation coefficients.
for (j in 1:length(R))
{
  r = R[j]

  ## Generate nrep sample correlation coefficients.
  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)
  }

  ## Generate nrep approximate 95% CI's

  logterm = 0.5*log(1+C)-0.5*log(1-C)
  CL = logterm-2/sqrt(n-3)
  CU = logterm+2/sqrt(n-3)

```

```

LB = (exp(2*CL)-1)/(exp(2*CL)+1)
UB = (exp(2*CU)-1)/(exp(2*CU)+1)

## Find the proportion of CI's
## that do not contain 0

POWER = c(POWER, mean( (LB > 0) | (UB < 0) ))
}

## Look at the output
print(cbind(R, POWER))

```

Consider generating the data as before, but now dichotomize the data using the rule: $A_i = \mathcal{I}(X_i \geq c)$, $B_i = \mathcal{I}(Y_i \geq c)$. Write a function in R that takes three arguments (X , Y , and c), where X and Y are n -length vectors. The function returns a 2x2 contingency table of the dichotomized data.

```

dichot=function(X,Y,c)
{
  ## Initialize the table
  N=array(0,c(2,2))

  ## Dichotomize the data
  A=1*(X>=c)
  B=1*(Y>=c)

  ## Populate the table
  N[1,1]=sum( (A==1) & (B==1) )
  N[1,2]=sum( (A==1) & (B==0) )
  N[2,1]=sum( (A==0) & (B==1) )
  N[2,2]=sum( (A==0) & (B==0) )
  return(N)
}

```

To exemplify the use of this function, let's generate `nrep` sample correlations as well as `nrep` sample log-odds ratios and standard errors from a dichotomized version of the data with $c = 0$. Consider $r = 0.01$.

```

n=100
r=0.01
## Generate nrep sample correlation coefficients.
C = array(0, nrep)
SLOR = array(0, nrep)

```

```
SE = 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)
  N=dichot(X,Y,c=0)
  SLOR[i] = log(N[1,1]) + log(N[2,2]) - log(N[1,2]) - log(N[2,1])
  SE[i] = sqrt(1/N[1,1] + 1/N[1,2] + 1/N[2,1] + 1/N[2,2])
}
```