# HW3: Bootstrap (Due: Thursday, Nov 10, at noon)

The following function simulates data from a bi-variate distribution.

```
simXY=function(n,rho){
    x=scale(rexp(n))
    y=x*rho+rnorm(n,sd=sqrt(1-rho^2))
    return(cbind(x,y))
}

# testing it
 tmp=simXY(1e6,.23)
 cor(tmp)
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.2308018
## [2,] 0.2308018 1.0000000
```

## 1) SE of the sample correlation and an approximate 95% CI

The following formula is commonly used to approximate the SE of the sample correlation

$SE = \sqrt{\frac{(1-\rho^2)}{(n-2)}}$

**1.1)** Using the following data set report the sample correlation, the SE and an approximate 95% CI (assuming normality) using the formula presented above.

```
 set.seed(195021)
 DATA_30=simXY(n=30,rho=.5)
```

```
  COR_30=cor(DATA_30[,1],DATA_30[,2])
  SE_30=sqrt((1-COR_30^2)/(nrow(DATA_30)-2))
  CI_30=COR_30+c(-1.96,1.96)*SE_30
  ANS=data.frame('COR' = round(COR_30,4), 'SE' = round(SE_30,4), 'LOWER' = round(CI_30[1],4), 'UPPER' =
  ANS
```

```
##     COR     SE  LOWER  UPPER
## 1 0.493 0.1644 0.1708 0.8153
```

**1.2)** Repeat 1.1 using a sample size of 300, comment on the differences in the results

```
 set.seed(195021)
 DATA_300=simXY(n=300,rho=.5)
```

```
  COR_300=cor(DATA_300[,1],DATA_300[,2])
  SE_300=sqrt((1-COR_300^2)/(nrow(DATA_300)-2))
  CI_300=COR_300+c(-1.96,1.96)*SE_300
  ANS=data.frame('COR' = round(COR_300,4), 'SE' = round(SE_300,4), 'LOWER' = round(CI_300[1],4), 'UPPER
  ANS
```

```
##     COR     SE  LOWER  UPPER
## 1 0.516 0.0496 0.4187 0.6132
```

## 2) Bootstrap CIs (percentile method)

Use 5000 Bootstrap samples to estimate the SE of the sample correlation, and an approximate 95% CI, for each of the data sets simulated above (DATA_30 and DATA_300).

To estimate the CI use the percentile method we used in class. That is, report the empirical 2.5% and 97.5% percentiles of the bootstrap estimates.

**Note:** as you implement bootstrap, be sure to save the bootstrap estimates in a vector, you will need those for questions 3 and 4 as well.

Compare your bootstrap CIs with the ones previously reported.

```
COR.B_30=rep(NA,5000)

for(i in 1:5000){
  tmp=sample(1:30,size=30,replace=TRUE)
  COR.B_30[i]=cor(DATA_30[tmp,])[2,1]
}
CI.B_30= quantile(COR.B_30,prob=c(.025,.975))
round(rbind('Conventional'=CI_30,'Percentile'=CI.B_30),4)
```

```
##                2.5%  97.5%
## Conventional 0.1708 0.8153
## Percentile   0.1044 0.7418
```

```
COR.B_300=rep(NA,5000)

for(i in 1:5000){
  tmp=sample(1:300,size=300,replace=TRUE)
  COR.B_300[i]=cor(DATA_300[tmp,])[2,1]
}
CI.B_300= quantile(COR.B_300,prob=c(.025,.975))
round(rbind('Conventional'=CI_300,'Percentile'=CI.B_300),4)
```

```
##                2.5%  97.5%
## Conventional 0.4187 0.6132
## Percentile   0.4159 0.6036
```

## 3) Bootstrap CI: pivotal method

An alternative approach for estimating bootstrap CI is as follows

- Collect bootstrap estimates $[r_1, r_2, ..., r_k]$, here $r_*$ is a bootstrap estimate of the correlation
- Subtract from the bootstrap estimate the mean of the bootstrap estimates $(\bar{r})$, that is: $[\tilde{r}_1 = (r_1 - \bar{r}), \tilde{r}_2 = (r_2 - \bar{r}), ..., \tilde{r}_k = (r_k - \bar{r})]$
- Compute the relevant percentiles (e.g., $q_{0.025}$, and $q_{0.975}$) of the $\tilde{r}$'s
- Use $CI_{95\%} = [r + q_{0.025}; r + q_{0.975}]$, where $r$ is the sample correlation evaluated in the original data set.

Report 95% pivotal CIs for the DATA_30 and DATA_300 using the method described above.

```
tmp_r=COR.B_30-mean(COR.B_30)
CI.B_30.2=COR_30+quantile(tmp_r,prob=c(0.025,.975))

round(rbind('Conventional'=CI_30,'Percentile'=CI.B_30,'Pivotal'=CI.B_30.2),4)
```

```
##                2.5%  97.5%
## Conventional 0.1708 0.8153
## Percentile   0.1044 0.7418
## Pivotal      0.1256 0.7630
```

```
tmp_r=COR.B_300-mean(COR.B_300)
CI.B_300.2=COR_300+quantile(tmp_r,prob=c(0.025,.975))

round(rbind('Conventional'=CI_300,'Percentile'=CI.B_300,'Pivotal'=CI.B_300.2),4)
```

```
##                2.5%  97.5%
## Conventional 0.4187 0.6132
## Percentile   0.4159 0.6036
## Pivotal      0.4185 0.6063
```

## 4) Bootstrap CI: normal method

If we assume normality, we can compute a Bootstrap CI using $r+/-1.96\times SE$ where $r$ is the correlation estimated in the original sample, and $SE$ is a Bootstrap estimate of the SE.

Report 95% CIs for each of the data sets using the normal method.

```
CI.B_30.3=COR_30+c(-1.96,1.96)*sd(COR.B_30)

round(rbind('Conventional'=CI_30,'Percentile'=CI.B_30,'Pivotal'=CI.B_30.2,'Normal'=CI.B_30.3),4)
```

```
##                2.5%  97.5%
## Conventional 0.1708 0.8153
## Percentile   0.1044 0.7418
## Pivotal      0.1256 0.7630
## Normal       0.1698 0.8163
```

```
CI.B_300.3=COR_300+c(-1.96,1.96)*sd(COR.B_300)

round(rbind('Conventional'=CI_300,'Percentile'=CI.B_300,'Pivotal'=CI.B_300.2,'Normal'=CI.B_300.3),4)
```

```
##                2.5%  97.5%
## Conventional 0.4187 0.6132
## Percentile   0.4159 0.6036
## Pivotal      0.4185 0.6063
## Normal       0.4227 0.6092
```