Homework 3

2022-11-07

```
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 0.2301564
## [2,] 0.2301564 1.0000000
```

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

1.1)

```
set.seed(195021)
DATA_30=simXY(n=30,rho=.5)
pho=cor(DATA_30)[1, 2]
SE=sqrt((1-pho**2)/(30-2))
lower=pho-1.96*SE
upper=pho+1.96*SE
CI_small_1=c(lower, upper)
cat("Sample Correlation: ", pho, "\n")
```

```
## Sample Correlation:  0.4930354
```

```
cat("The SE: ", SE, "\n")
```

```
## The SE:  0.1644163
```

```
cat("Approximate 95% CI (assuming normality): ", CI_small_1)
```

```
## Approximate 95% CI (assuming normality):  0.1707795 0.8152913
```

1.2)

```
set.seed(195021)
DATA_300=simXY(n=300,rho=.5)
pho=cor(DATA_300)[1, 2]
SE=sqrt((1-pho**2)/(300-2))
lower=pho-1.96*SE
upper=pho+1.96*SE
CI_large_1=c(lower, upper)
cat("Sample Correlation: ", pho, "\n")
```

```
## Sample Correlation:  0.5159531
```

```
cat("The SE: ", SE, "\n")
```

```
## The SE:  0.04962248
```

```
cat("Approximate 95% CI (assuming normality): ", CI_large_1)
```

```
## Approximate 95% CI (assuming normality):  0.418693 0.6132131
```

Since the number of samples becomes larger, the SE becomes smaller. Thus, the 95% CI has become smaller. The Sample Correlation is similar to the small dataset.

2) Bootstrap CIs (percentile method)

For small dataset:

```
set.seed(195021)
SEs_small = rep(1, 5000)
phos_small = rep(1, 5000)
for (i in 1:5000){
  index = sample(30, replace = TRUE)
  current = DATA_30[index, ]
  phos_small[i]=cor(current)[1, 2]
  SEs_small[i]=sqrt((1-phos_small[i]**2)/(30-2))
}
lower=quantile(phos_small, 0.025)
upper=quantile(phos_small, 0.975)
CI_small_2=c(lower, upper)
cat("Sample Correlation: ", mean(phos_small), "\n")
```

```
## Sample Correlation:  0.4720959
```

```
cat("The SE: ", mean(SEs_small), "\n")
```

```
## The SE:  0.1628254
```

```
cat("Approximate 95% CI (Bootstrap CIs percentile method): ", CI_small_2)
```

```
## Approximate 95% CI (Bootstrap CIs percentile method):  0.09457221 0.7349033
```

For large dataset:

```
set.seed(195021)
SEs_large = rep(1, 5000)
phos_large = rep(1, 5000)
for (i in 1:5000){
  index = sample(300, replace = TRUE)
  current = DATA_300[index, ]
  phos_large[i]=cor(current)[1, 2]
  SEs_large[i]=sqrt((1-phos_large[i]**2)/(300-2))
}
lower=quantile(phos_large, 0.025)
upper=quantile(phos_large, 0.975)
CI_large_2=c(lower, upper)
cat("Sample Correlation: ", mean(phos_large), "\n")
```

```
## Sample Correlation:  0.513369
```

```
cat("The SE: ", mean(SEs_large), "\n")
```

```
## The SE:  0.04960615
```

```
cat("Approximate 95% CI (Bootstrap CIs percentile method): ", CI_large_2)
```

```
## Approximate 95% CI (Bootstrap CIs percentile method):  0.4141608 0.6045546
```

3) Bootstrap CI: pivotal method

For small dataset:

```
r_avg = mean(phos_small)
r_minus_avg = phos_small - r_avg
r_minus_avg = r_minus_avg[order(r_minus_avg)]
lower=cor(DATA_30)[1, 2]+quantile(r_minus_avg, 0.025)
upper=cor(DATA_30)[1, 2]+quantile(r_minus_avg, 0.975)
CI_small_3=c(lower, upper)
cat("Approximate 95% CI (Bootstrap CIs pivotal method): ", CI_small_3)
```

```
## Approximate 95% CI (Bootstrap CIs pivotal method):  0.1155118 0.7558429
```

For large dataset:

```
r_avg = mean(phos_large)
r_minus_avg = phos_large - r_avg
r_minus_avg = r_minus_avg[order(r_minus_avg)]
lower=cor(DATA_300)[1, 2]+quantile(r_minus_avg, 0.025)
upper=cor(DATA_300)[1, 2]+quantile(r_minus_avg, 0.975)
CI_large_3=c(lower, upper)
cat("Approximate 95% CI (Bootstrap CIs pivotal method): ", CI_large_3)
```

```
## Approximate 95% CI (Bootstrap CIs pivotal method):  0.4167449 0.6071386
```

4) Bootstrap CI: normal method

For small dataset:

```
lower=cor(DATA_30)[1, 2]-1.96*mean(SEs_small)
upper=cor(DATA_30)[1, 2]+1.96*mean(SEs_small)
CI_small_4=c(lower, upper)
cat("Approximate 95% CI (Bootstrap CIs normal method): ", CI_small_4)
```

```
## Approximate 95% CI (Bootstrap CIs normal method):  0.1738977 0.8121731
```

For large dataset:

```
lower=cor(DATA_300)[1, 2]-1.96*mean(SEs_large)
upper=cor(DATA_300)[1, 2]+1.96*mean(SEs_large)
CI_large_4=c(lower, upper)
cat("Approximate 95% CI (Bootstrap CIs normal method): ", CI_large_4)
```

```
## Approximate 95% CI (Bootstrap CIs normal method):  0.418725 0.6131811
```

5) Compare all the CIs for each of the datasets:

For small dataset:

```
res_small=rbind(CI_small_1, CI_small_2, CI_small_3, CI_small_4)
rownames(res_small)=c('CIs assuming normality',
                      'Bootstrap CIs percentile method',
                      'Bootstrap CIs pivotal method',
                      'Bootstrap CIs normal method')
res_small
```

```
##                                      2.5%     97.5%
## CIs assuming normality          0.17077948 0.8152913
## Bootstrap CIs percentile method 0.09457221 0.7349033
## Bootstrap CIs pivotal method    0.11551177 0.7558429
## Bootstrap CIs normal method     0.17389769 0.8121731
```

For large dataset:

```
res_large=rbind(CI_large_1, CI_large_2, CI_large_3, CI_large_4)
rownames(res_large)=c('CIs assuming normality',
                      'Bootstrap CIs percentile method',
                      'Bootstrap CIs pivotal method',
                      'Bootstrap CIs normal method')
res_large
```

```
##                                      2.5%     97.5%
## CIs assuming normality          0.4186930 0.6132131
## Bootstrap CIs percentile method 0.4141608 0.6045546
## Bootstrap CIs pivotal method    0.4167449 0.6071386
## Bootstrap CIs normal method     0.4187250 0.6131811
```

For the small dataset, the results obtained by the four methods are quite different. But for the large dataset, the results obtained by the four methods are not much different. This shows that the number of samples affects the statistical results.