

Statistics 406 Problem Set 5

Due in lab, Tuesday October 23

1. Suppose we observe iid values X_1, \dots, X_n that are uniformly distributed on the interval $(0, a)$, where $a > 0$ is an unknown constant. We can estimate a using the maximum value of the sample:

$$\hat{a} = \max(X_1, \dots, X_n).$$

Since $\hat{a} < a$ it has negative bias. The relative bias is

$$\frac{E\hat{a} - a}{a}$$

It is a fact that the relative bias in this setting is $-1/(n+1)$.

- (a) Use simulation with at least three different sample sizes to determine how the non-parametric bootstrap performs at estimating the relative bias.

Solution:

Here is the code:

```
## Number of simulation replications.
nrep = 100

## Number of bootstrap samples.
nboot = 1000

## Population value of the upper limit of the uniform distribution.
a = 2

## Sample sizes.
SS = c(5,10,20)

## Storage for the bias estimates.
Bias = NULL

## Loop over the sample sizes.
for (k in 1:3) {

  ## The sample size for the current iteration.
  n = SS[k]
```

```

bias = NULL
for (r in 1:nrep) {

  ## Generate a sample from the uniform population.
  X = runif(n, max=a)

  ## Generate non-parametric bootstrap samples.
  ii = ceiling(n*runif(nboot*n))
  Xboot = X[ii]
  Xboot = array(Xboot, c(nboot, n))

  ## The bootstrap estimate of the relative bias.
  MX = apply(Xboot, 1, max)
  bias[r] = (mean(MX) - max(X))/max(X)
}

## The overall estimate of the relative bias.
Bias[k] = mean(bias)
}

```

Here is what I got:

```

> Bias
[1] -0.07985626 -0.05047077 -0.02602383
> -1/(SS+1)
[1] -0.16666667 -0.09090909 -0.04761905

```

The non-parametric bootstrap doesn't seem to work very well here.

- (b) Use simulation with the same three sample sizes as in part (a) to determine how the parametric bootstrap (assuming a uniform distribution) performs at estimating the relative bias.

Solution:

```

## Number of simulation replications.
nrep = 100

## Number of bootstrap samples.
nboot = 1000

```

```

## Upper limit of the uniform population.
a = 2

## Sample sizes.
SS = c(5,10,20)

## Loop over sample sizes.
Bias = NULL
for (k in 1:3) {

  ## The sample size for the current iteration.
  n = SS[k]

  ## Simulation loop.
  bias = NULL
  for (r in 1:nrep) {

    ## Generate data from the uniform population.
    X = runif(n, max=a)

    ## Generate parametric bootstrap samples.
    mx = max(X)
    Xboot = runif(nboot*n, max=mx)
    Xboot = array(Xboot, c(nboot, n))

    ## The bootstrap estimate of the relative bias.
    MX = apply(Xboot, 1, max)
    bias[r] = (mean(MX) - mx)/mx
  }

  ## The overall bias estimate.
  Bias[k] = mean(bias)
}

```

The parametric bootstrap works very well:

```

> Bias
[1] -0.16697932 -0.09085892 -0.04757065
> -1/(SS+1)
[1] -0.16666667 -0.09090909 -0.04761905

```

2. Suppose we have samples from two populations: X_1, \dots, X_{n_x} are iid from population 1, and Y_1, \dots, Y_{n_y} are iid from population 2. Our goal is to estimate EX/EY ,

(a) Suppose we decide to use \bar{X}/\bar{Y} as our estimator. Use simulation with at least three sample sizes to estimate the bias, variance, and MSE of this estimator.

Solution:

Here is my code:

```
## Sample sizes.
SS = c(10,20,30)

## We'll use Gaussian populations with these expected values and
## variances.
EX = 1
EY = 2
SDX = 1
SDY = 1

## Number of simulation replications.
nrep = 1e4

bias = NULL
variance = NULL
MSE = NULL

## Loop over the sample sizes.
for (k in 1:3) {

  ## The sample size for this iteration.
  n = SS[k]

  ## Generate data from the X population.
  X = rnorm(nrep*n, mean=EX, sd=SDX)
  X = array(X, c(nrep,n))

  ## Generate data from the Y population.
  Y = rnorm(nrep*n, mean=EY, sd=SDY)
  Y = array(Y, c(nrep,n))

  ## The plug-in estimates of EX/EY.
  Xbar = apply(X, 1, mean)
```

```

Ybar = apply(Y, 1, mean)
R = Xbar/Ybar

## Calculate bias, variance, and MSE.
bias[k] = mean(R) - EX/EY
variance[k] = var(R)
MSE[k] = mean( (R-EX/EY)^2 )
}

```

Here is what I got:

```

> bias
[1] 0.013961868 0.005298828 0.005543078
> variance
[1] 0.03497678 0.01655705 0.01097354
> MSE
[1] 0.03516822 0.01658348 0.01100317

```

(b) Now suppose we decide to use

$$\bar{X}/\bar{Y} - \bar{X}\hat{\sigma}_Y^2/(n_y\bar{Y}^3)$$

as our estimator, where $\hat{\sigma}_Y^2$ is the sample variance of the Y_i . Use simulation with the same three sample sizes to estimate the bias, variance, and MSE of this estimator. Compare what happens in this situation to part (a).

Solution:

Here is my code:

```

## Sample sizes.
SS = c(10,20,30)

## We'll use Gaussian populations with these expected values and variances.
EX = 1
EY = 2
SDX = 1
SDY = 1

## Number of simulation replications.
nrep = 1e4

bias = NULL

```

```

variance = NULL
MSE = NULL

for (k in 1:3) {

  ## The sample size for this iteration.
  n = SS[k]

  ## Simulate data from the X population.
  X = rnorm(nrep*n, mean=EX, sd=SDX)
  X = array(X, c(nrep,n))

  ## Simulate data from the Y population.
  Y = rnorm(nrep*n, mean=EY, sd=SDY)
  Y = array(Y, c(nrep,n))

  ## The estimate of EX/EY.
  Xbar = apply(X, 1, mean)
  Ybar = apply(Y, 1, mean)
  VY = apply(Y, 1, var)
  R = Xbar/Ybar - Xbar*VY/(n*Ybar^3)

  ## Calculate the bias, variance, and MSE.
  bias[k] = mean(R) - EX/EY
  variance[k] = var(R)
  MSE[k] = mean( (R-EX/EY)^2 )
}

```

Here is what I got:

```

> bias
[1] -0.0010246684 -0.0002944437  0.0002237040
> variance
[1] 0.03281010 0.01561942 0.01043643
> MSE
[1] 0.03280787 0.01561795 0.01043544

```

Compared to the plug-in estimate, this estimate has very little bias. The variance is similar to the variance of the plug-in estimate. Since the bias is reduced and the variance is unchanged, the MSE is reduced. Since the variance contributes much more than the squared bias in this problem, the reduction in MSE is slight.

3. Suppose we observe data from a contaminated normal distribution. As described in detail in the notes, fraction $1 - p$ of the data are standard normal, which is the distribution whose parameters we intend to estimate. The complementary fraction p of the data come from a normal distribution with mean zero and standard deviation 10. We are interested in both the standard deviation and interquartile range (IQR) for the non-outlier part of the data (so the population values are $\text{IQR} = 1.35$ and $\sigma = 1$). Use simulation to estimate the relative RMSE (i.e. the RMSE divided by the estimation target) for these estimators based on contaminated normal data as described above with $p = 0, 0.05, 0.1, 0.15,$ and 0.2 . Briefly describe how these two estimators are affected by the presence of outliers.

Solution:

```
## Fraction of contaminated data.
P = c(0, 0.05, 0.1, 0.15, 0.2)

## Sample size.
n = 20

## Number of simulation replications.
nrep = 1e4

M = array(0, c(length(P), 3))

## Loop over the contamination fractions.
for (k in 1:length(P)) {

  X = rnorm(nrep*n)
  X = array(X, c(nrep,n))

  Y = rnorm(nrep*n, sd=10)
  Y = array(Y, c(nrep,n))

  A = 1*(runif(nrep*n) < P[k])
  A = array(A, c(nrep,n))

  X = (1-A)*X + A*Y

  ## Calculate the IQR and SD.
  I = apply(X, 1, IQR)
  S = apply(X, 1, sd)

  ## Get the relative RMSE's.
```

```
ei = sqrt(mean( (I - 1.35)**2 )) / 1.35
es = sqrt(mean( (S - 1)**2 ))

M[k,] = c(P[k], ei, es)
}
```

I get this:

```
> M
      [,1]      [,2]      [,3]
[1,] 0.00 0.2459998 0.1614920
[2,] 0.05 0.2576837 1.7295668
[3,] 0.10 0.2929022 2.4517774
[4,] 0.15 0.3765341 3.1193324
[5,] 0.20 0.5247593 3.6627026
```

This shows that both the IQR and $\hat{\sigma}$ are biased substantially upward by the presence of outliers, but $\hat{\sigma}$ is much more biased than IQR.