STAT 601 Homework 3

Yifan Zhu

February 11, 2018

For each $M \in \{100, 200, \ldots, 1000, 2000\}$, we simulated x_m, y_m, $m = 1, 2, \ldots, M$ iid from $N(0,1)$ for N times. Each time we compute $P_{1,M} = \frac{1}{M}\sum_{m=1}^{M} I(x_m \leq y_m)$ and $P_{2,M} = \frac{1}{M^2}\sum_{j=1}^{M}\sum_{m=1}^{M} I(x_m \leq y_j)$. Then for each M, we have $P_{1,M,i}, P_{2,M,i}$, $i = 1, 2, \ldots, N$. Hence for each M, the approxtimated $\mathrm{E}(P_{t,M})$ and $\mathrm{Var}(P_{t,M})$ are

$$\widehat{\mathrm{E}}(P_{t,M}) = \sum_{i=1}^{N} P_{t,M,i}$$

and

$$\widehat{\mathrm{Var}}(P_{t,M}) = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_{t,M,i} - \widehat{\mathrm{E}}(P_{t,M})\right)^2,$$

where $t = 1, 2$. Because x_m, y_m are iid from $N(0,1)$, then the true value is $\mathrm{E}(X \leq Y) = 1/2$. And the results we got are shown below.

For $t = 1$, $\widehat{\mathrm{E}}(P_{1,M})$ for each M are

```
0.5024250 0.5005275 0.5010450 0.4994500 0.4998200 0.5001033 0.4995036
 ↪ 0.5006388 0.4997372 0.5002655 0.5000510
```

and $\widehat{\mathrm{Var}}(P_{1,M})$ for each M are

```
0.0025018203 0.0012156671 0.0007939994 0.0006500538 0.0005076214
 ↪ 0.0004035550 0.0003632505 0.0003204710
```

For $t = 2$, $\widehat{\mathrm{E}}(P_{2,M})$ for each M are

```
0.5017982 0.5005340 0.5008111 0.4996622 0.4997396 0.4999114 0.4997565
 ↪ 0.5003425 0.4998615 0.5002281 0.4998734
```

and $\widehat{\mathrm{Var}}(P_{2,M})$ for each M are

```
1.697511e-03 8.190077e-04 5.306455e-04 4.245553e-04 3.452576e-04
 ↪ 2.816254e-04 2.580459e-04 2.180506e-04 1.805410e-04 1.697915e-04
 ↪ 8.446357e-05
```

And the ratio of approximated variances $\widehat{\mathrm{Var}}(P_{1,M})/\widehat{\mathrm{Var}}(P_{2,M})$ for each M are

```
1.473817 1.484317 1.496290 1.531140 1.470269 1.432950 1.407697
 ↪ 1.469709 1.502976 1.572620 1.459541
```

So we can see both methods gives an estimator of $\mathrm{E}(X \leq Y)$ with expected value around the true one (1/2). But the precition of $P_{2,M}$ seems to be higher, With smaller variance relative to $P_{1,M}$, since the ratio of variances is around 1.5.

Next we consider $P_M(1-P_M)/M$. Since

$$\widehat{\mathrm{Var}}(P_{t,M}) = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_{t,M,i} - \widehat{\mathrm{E}}(P_{t,M})\right)^2 = \frac{N}{N-1}\left(\frac{\sum_{i=1}^{N} P_{t,Mi}^2}{N} - \left(\frac{\sum_{i=1}^{N} P_{t,M,i}}{N}\right)^2\right),$$

which approximates $\mathrm{E}(P_{t,M}^2) - (\mathrm{E}(P_{t,M}))^2 = \mathrm{Var}(P_{t,M})$ as $N \to \infty$. So we want to compare $\mathrm{Var}(P_{t,M}) = \mathrm{E}(P_{t,M}^2) - (\mathrm{E}(P_{t,M}))^2$ and $\mathrm{E}\left(\frac{P_{t,M}(1-P_{t,M})}{M}\right) = \frac{1}{M}\left(\mathrm{E}(P_{t,M}) - \mathrm{E}(P_{t,M}^2)\right)$.

Let $p = P(X \le Y)$. Then $E(I(x_i \le y_j)) = p$. When $t = 1$,

$$\mathrm{E}(P_{1,M}) = \mathrm{E}\left(\frac{1}{M}\sum_{m=1}^{M} I(x_m \le y_m)\right) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{E}(I(x_m \le y_m)) = \frac{1}{M}Mp = p$$

and we note that $MP_{1,M} = \sum_{m=1}^{M} I(x_m \le y_m)$ is sum of iid Bernolli r.v's, so $MP_{1,M} \sim \mathrm{Binomial}(M,p)$. Hence

$$\mathrm{E}(P_{1,M}^2) = \frac{1}{M^2}\mathrm{E}\left(M^2P_{1,M}^2\right) = \frac{1}{M^2}(\mathrm{Var}(MP_{1,M}) + (\mathrm{E}(MP_{1,M}))^2) = \frac{1}{M^2}(Mp(1-p) + M^2p^2) = \frac{p(1-p)}{M} + p^2$$

Then

$$\mathrm{Var}(P_{1.M}) = \frac{p(1-p)}{M} + p^2 - p^2 = \frac{p(1-p)}{M}$$

and

$$\mathrm{E}\left(\frac{P_{1,M}(1-P_{1,M})}{M}\right) = \frac{1}{M}\left(p - \frac{p(1-p)}{M} - p^2\right) = \frac{p(1-p)}{M} - \frac{p(1-p)}{M^2}$$

So the difference is in $p(1-p)/M^2$ and when M is large they are close.

When $t = 2$, $P_{2,M} = \frac{1}{M^2}\sum_{i,j=1}^{M} I(x_i \le y_j)$, so

$$\mathrm{E}(P_{2,M}) = \frac{1}{M^2}\sum_{i,j=1}^{M}\mathrm{E}(I(x_i \le y_j)) = \frac{1}{M^2}M^2p = p$$

and

$$\begin{aligned}
\mathrm{E}(P_{2,M}^2) &= \mathrm{E}\left(\frac{1}{M^4}\left(\sum_{i,j} I(x_i \le y_j)\right)^2\right)\\
&= \frac{1}{M^4}\mathrm{E}(\sum_{i,j,k,l} I(x_i \le y_j)I(x_k \le y_l))\\
&= \frac{1}{M^4}\sum_{i,j,k,l}\mathrm{E}(I(x_i \le y_j)I(x_k \le y_l))
\end{aligned}$$

If $i = k$, $j = l$,

$$\mathrm{E}(I(x_i \le y_j)I(x_k \le y_l)) = \mathrm{E}\left((I(x_i \le y_j))^2\right) = \mathrm{E}(I(x_i \le y_j)) = p$$

If $i = k$, $j \ne l$,

$$\begin{aligned}
&\mathrm{E}(I(x_i \le y_j)I(x_k \le y_l))\\
=&\mathrm{E}(I(x_i \le y_j)I(x_i \le y_l))\\
=&\mathrm{E}\left[\mathrm{E}\left[I(y_j \ge x_i)I(y_l \ge x_i)|x_i\right]\right]\\
=&\mathrm{E}\left[\mathrm{E}[I(y_j \ge x_i)|x_i]\,\mathrm{E}[I(y_l \ge x_i)|x_i]\right]\\
=&\mathrm{E}[(1-F_Y(x_i))^2]\\
=&C_1
\end{aligned}$$

If $i \neq k$, $j = l$,

$$\begin{aligned}
& \mathrm{E}(I(x_i \leq y_j) I(x_k \leq y_l)) \\
=& \mathrm{E}(I(x_i \leq y_j) I(x_k \leq y_j)) \\
=& \mathrm{E}\left[\mathrm{E}\left[I(x_i \leq y_j) I(x_k \leq y_j) | y_j\right]\right] \\
=& \mathrm{E}\left[\mathrm{E}[I(x_i \leq y_j) | y_j]\, \mathrm{E}[I(x_k \leq y_j) | y_j]\right] \\
=& \mathrm{E}[(F_X(y_j))^2] \\
=& C_2
\end{aligned}$$

If $i \neq k$, $j \neq l$,

$$\mathrm{E}(I(x_i \leq y_j) I(x_k \leq y_l) = \mathrm{E}(x_i \leq y_j)\, \mathrm{E}(x_k \leq y_l) = p^2$$

Hence

$$\begin{aligned}
& \sum_{i,j,k,l} \mathrm{E}(I(x_i \leq y_j) I(x_k \leq y_l)) \\
=& M(M-1) \cdot M(M-1) \cdot p^2 + M \cdot M \cdot p + M(M-1) \cdot M \cdot C_1 + M \cdot M(M-1) \cdot C_2 \\
=& M^2(M-1)^2 p^2 + M^2 p + M^2(M-1)(C_1 + C_2)
\end{aligned}$$

Thus

$$\mathrm{E}(P_{2,M}^2) = \left(\frac{M-1}{M}\right)^2 p^2 + \frac{p}{M^2} + \frac{M-1}{M^2}(C_1 + C_2)$$

Then

$$\mathrm{Var}(P_{2,M}) = \left(\frac{M-1}{M}\right)^2 p^2 + \frac{p}{M^2} + \frac{M-1}{M^2}(C_1 + C_2) - p^2 = \frac{p}{M^2} - \frac{2M-1}{M^2} p^2 + \frac{M-1}{M^2}(C_1 + C_2)$$

and

$$\begin{aligned}
& \mathrm{E}\left(\frac{P_{2,M}(1 - P_{2,M})}{M}\right) \\
=& \frac{1}{M}\left\{p - \left(\frac{M-1}{M}\right)^2 p^2 - \frac{p}{M^2} - \frac{M-1}{M^2}(C_1 + C_2)\right\} \\
=& \frac{M^2-1}{M^3} p - \frac{(M-1)^2}{M^3} p^2 - \frac{M-1}{M^3}(C_1 + C_2)
\end{aligned}$$

So they are different.

And we can also show when X and Y are from the same continuous distribution, $F_X = F_y$, $F_X(y_i) = F_Y(x_i)$ and $1 - F_Y(x_i) = 1 - F_X(x_i)$ are Uniform$(0,1)$. Thus $C_1 = C_2 = 1/3$, and $p = \mathrm{E}(X \leq Y) = 1/2$. Then

$$\mathrm{Var}(P_{1,M}) = \frac{1}{4M}$$

$$\mathrm{Var}(P_{2,M}) = \frac{1}{2M^2} - \frac{2M-1}{4M^2} + \frac{2(M-1)}{3M^2} = \frac{2M+1}{12M^2}$$

The ratio is then

$$\frac{\mathrm{Var}(P_{1,M})}{\mathrm{Var}(P_{2,M})} = \frac{1}{4M} \frac{12M^2}{2M+1} = \frac{3M}{2M+1}$$

This ratio is about 1.5 when M is large, which jusutifies the simulation results we got.

R Code

```r
Ms <- c(100*1:10, 2000)
N <- 2000

P1Ms <- rep(0, N)
P2Ms <- P1Ms

P1M <- function(x,y){
  M <- length(x)
  return(sum(x<y)/M)
}

P2M <- function(x,y){
  M <- length(x)
  x1 <- rep(x, times = M)
  y1 <- rep(y, each = M)
  return(sum(x1 < y1)/M^2)
}
P1Mss <- NULL
P2Mss <- NULL
for (M in Ms){
  xs <- matrix(rnorm(M*N), ncol = N)
  ys <- matrix(rnorm(M*N), ncol = N)
  for (n in 1:N){
    P1Ms[n] <- P1M(xs[,n],ys[,n])
    P2Ms[n] <- P2M(xs[,n],ys[,n])
  }
  P1Mss <- cbind(P1Mss, P1Ms)
  P2Mss <- cbind(P2Mss, P2Ms)
}
ep1m <- apply(P1Mss, MARGIN = 2, FUN = mean)
ep2m <- apply(P2Mss, MARGIN = 2, FUN = mean)
varp1m <- apply(P1Mss, MARGIN = 2, FUN = var)
varp2m <- apply(P2Mss, MARGIN = 2, FUN = var)

result <- list(ep1m = ep1m, ep2m = ep2m, varp1m = varp1m, varp2m =
 ↪  varp2m)

save(P1Mss, P2Mss, result, file = "./results.Rda")
```