

Sampling_HW3

김민국

2019-11-5

Problem3

```
x <- c(3,5,8,8,12)
y <- c(0,0,2,3,3)
mu_x <- mean(x) ## x의 모평균
mu_y <- mean(y) ## y의 모평균

set.seed(10)
results <- c()
for(i in 1:100000){
  a <- sample(1:5,3) ## 3개의 sample 추출
  y_lm <- lm(y[a]~x[a]) ## sample을 바탕으로 얻은 회귀식
  y_lm_coefficients <- y_lm$coefficients
  y_hat <- y_lm_coefficients[1] + y_lm_coefficients[2] * mu_x ## y의 선형회귀 추정량
  results[i] <- y_hat
}
all.equal(mu_y, mean(results))
```

```
## [1] "Mean relative difference: 0.07738727"
```

-> 모집단에서 크기가 3인 sample을 추출하여 회귀추정량을 구하는 과정이다.

-> 회귀추정량의 기대값을 100000개의 회귀추정량의 평균으로 생각하였다.

-> 이 때 회귀추정량의 기대값과 우리가 알고 있는 모평균과 비교를 하였다.

```
set.seed(10)
r <- c()
for(i in 1:100000){
  a <- sample(1:5,3) ## 3개의 sample 추출
  y_bar <- mean(y[a]) ## y의 선형회귀 추정량
  r[i] <- y_bar
}
all.equal(mu_y, mean(r))
```

```
## [1] "Mean relative difference: 0.0002041667"
```

-> 대표적인 unbiased 추정량인 표본평균을 이용하여 비교해보았다.

-> 표본평균의 기대값 역시 회귀추정량과 마찬가지로 100000개의 표본평균의 평균으로 생각하였다.

-> 이렇게 얻은 기대값과 모평균을 비교해보았다.

-> 표본평균의 경우 Mean relative difference가 0.0002041667이고 회귀추정량의 경우 0.07738727

-> 표본평균이 대표적인 unbiased이므로 이를 기준으로 생각했을 때 회귀추정량은 표본평균에 비교하여 300배정도 차이난다고 볼 수 있다.

-> 따라서 회귀추정량은 biased하다고 생각할 수 있다.

Problem5

```
k <- seq(1:370) ## 전체 가구 번호
i <- c(28,31:33,36:41,44:45,47,55:56,58,68:69,82:83,85:86,89:94,98:99,101,107:110,
      114,154,156,178,223:224,296,298:300,302:304,306:323,325:331,333,335:339,341,342)
      ## 주택을 소유한 가구의 번호
kk <- rep(0,370) ## 주택 소유여부 확인하는 집합
kk[i] <- 1 ## 주택을 소유했으면 1, 소유하지 않았으면 0
p <- mean(kk) ## 모비율
```

```
set.seed(10)
r1 <- c() ## 계통표집을 통해 얻은 표본비율 값들
r2 <- c() ## SRS를 통해 얻은 표본비율 값들
for(i in 1:100000){
  a <- sample(1:8,1) ## 계통표집의 초기 시작점 임의로 설정
  b <- 1
  r <- c() ## 계통표집으로 추출되는 값들
  while(a < 363){
    r[b] <- kk[a]
    a <- a + 8
    b <- b + 1
  }
  r1[i] <- mean(r) ## 계통표집으로 얻은 표본비율
  s <- sample(1:370,b) ## 계통표집과 표본의 크기가 동일한 SRS
  r2[i] <- mean(kk[s]) ## srs를 통해 얻은 표본비율
}
```

```
r11 <- c() ## 계통표집의 MsE구하기 위한 결과
r21 <- c() ## 계통 표집의 MSE 구하기 위한 결과
for(i in 1:length(r1)){
  r11[i] <- (r1[i] - p)^2
  r21[i] <- (r2[i] - p)^2
}
mean(r11) ## 계통표집의 MSE
```

```
## [1] 0.001400565
```

```
mean(r21) ## SRS의 MSE
```

```
## [1] 0.003243629
```

- > 계통표집을 통해 얻은 표본비율의 MSE와 SRS를 통해 얻은 표본비율의 MSE를 비교해보았다.
- > 각 표집마다 반복수는 10000으로 통일하였으며 10000개의 평균을 기대값으로 생각하였다.
- > 둘의 효율성을 비교하기 위해 MSE를 이용하였다.
- > 100000개의 sample을 통해 MSE의 정의인 $E[(Y_{\text{hat}} - Y)^2]$ 를 각각 구했다.
- > 이렇게 구한 MSE를 비교한 결과 계통표집의 MSE가 SRS의 MSE보다 작았다.
- > 계통표집의 효율이 SRS보다 좋다고 볼 수 있다.