Midterm Election Simulation Project

Group 2

Erica Nam, Xiao Xu, Xinyu Zhang, Yitao Li

Data Simulation

For each of *nrep* replications:

- We first generate *N_contam* the size of the contaminating group using a bin(N, q). q is our assumption for the percentage of the contaminating group out of the entire sample which has a sample size N. N - *N_contam* is the size of the sample that is not contaminated. Note that when q = 0, there is no contamination in the sample.

- Second, we generate the number of people in the contaminating group who prefer candidate 1 (N_true_1), candidate 2 (N_true_2) and candidate 3 (N_true_3) respectively using a multinomial (*N_contam*, p_1, p_2, p_3) since the proportion of people in this group favors candidate 1, candidate 2 and candidate 3 is p_1, p_2, p_3.

- We also generate the number of people in the not contaminated group who prefer candidate 1 (N_contam_1), candidate 2 (N_contam_2) and candidate 3 (N_contam_3) respectively using a multinomial (N - *N_contam*, r_1, r_2, r_3) since the proportion of people in this group favors candidate 1, candidate 2 and candidate 3 is r_1, r_2, r_3.

- Then we have the simulated number of people who prefer candidate 1 $N_1 = N\_true_1 + N\_contam_1$. People who prefer candidate 2 $N_2 = N\_true_2 + N\_contam_2$. People who prefer candidate 3 $N_3 = N\_true_3 + N\_contam_3$.

Confidence Interval

Continue for the same replication *t*:

- We first calculate $p_1\_hat = \frac{N_1}{N}$. $p_2\_hat = \frac{N_2}{N}$. $p_3\_hat = \frac{N_3}{N}$.

- Then we calculate $CI\_m1_{i,j}$ the 1- α confidence interval for p_i - p_j using the following formula from method 1:

  $p_i\_hat - p_j\_hat \pm \sqrt{\frac{Ad_{i,j}}{N}}$ where $A = \chi^2_{M-1}\left(\frac{\alpha}{M}\right)$, $M = m\frac{(m-1)}{2}$, $m = 3, d_{i,j} = p_i\_hat + p_j\_hat - (p_i\_hat - p_j\_hat)^2$

  Calculate $CI\_m2_{i,j}$ the confidence interval for p_i - p_j using the following formula from method 2:

  $p_i\_hat - p_j\_hat \pm \frac{a}{\sqrt{N}}$ where$1 - 2[1 - z_a] - 4[m - 2][1 - z_{a\sqrt{2}}] = 1 - \alpha,\ m = 3$

  Note that if $p_1 - p_2 \in CI\_m11{,}2$ and $p_1 - p_3 \in CI\_m11{,}_3$ and $p_2 - p_3 \in CI\_m12{,}3$, then we say the simultaneous confidence interval consist of $CI\_m1_{1,2}$, *$CI\_m1_{1,3}$ and $CI\_m1_{2,3}$ covers the true value* $p_1 - p_2$, $p_1 - p_3$ *and* $p_2 - p_3$ simultaneously for this replication. Same goes for method 2.

- The average confidence interval width is $avg\_width\_m1_t = \frac{2\sqrt{\frac{Ad_{1,2}}{N}} + 2\sqrt{\frac{Ad_{1,3}}{N}} + 2\sqrt{\frac{Ad_{2,3}}{N}}}{3}$ for method 1 and $avg\_width\_m2_t = 2\frac{a}{\sqrt{N}}$ for method 2.

- The maximum confidence interval width is $max\_width\_m1_t = \max[\,2\sqrt{\frac{Ad_{1,2}}{N}}, 2\sqrt{\frac{Ad_{1,3}}{N}}, 2\sqrt{\frac{Ad_{2,3}}{N}}\,]$ for method 1 and $max\_width\_m2_t = 2\frac{a}{\sqrt{N}}$ for method 2.

Simultaneous IC coverage probability and width

- The CI and its width is calculated *nrep* times for *nrep* replications.

- The estimated coverage probability is $\frac{nrep\_cov\_m1}{nrep}$ if among *nrep* replications *nrep_cov_m1* of them cover the true value simultaneously with the CI constructed with method 1. Same foes for method 2.

- The estimated average simultaneous CI width is $\frac{\sum_t avg\_width\_m1_t}{nrep}$ for method 1. Same goes for method 2.

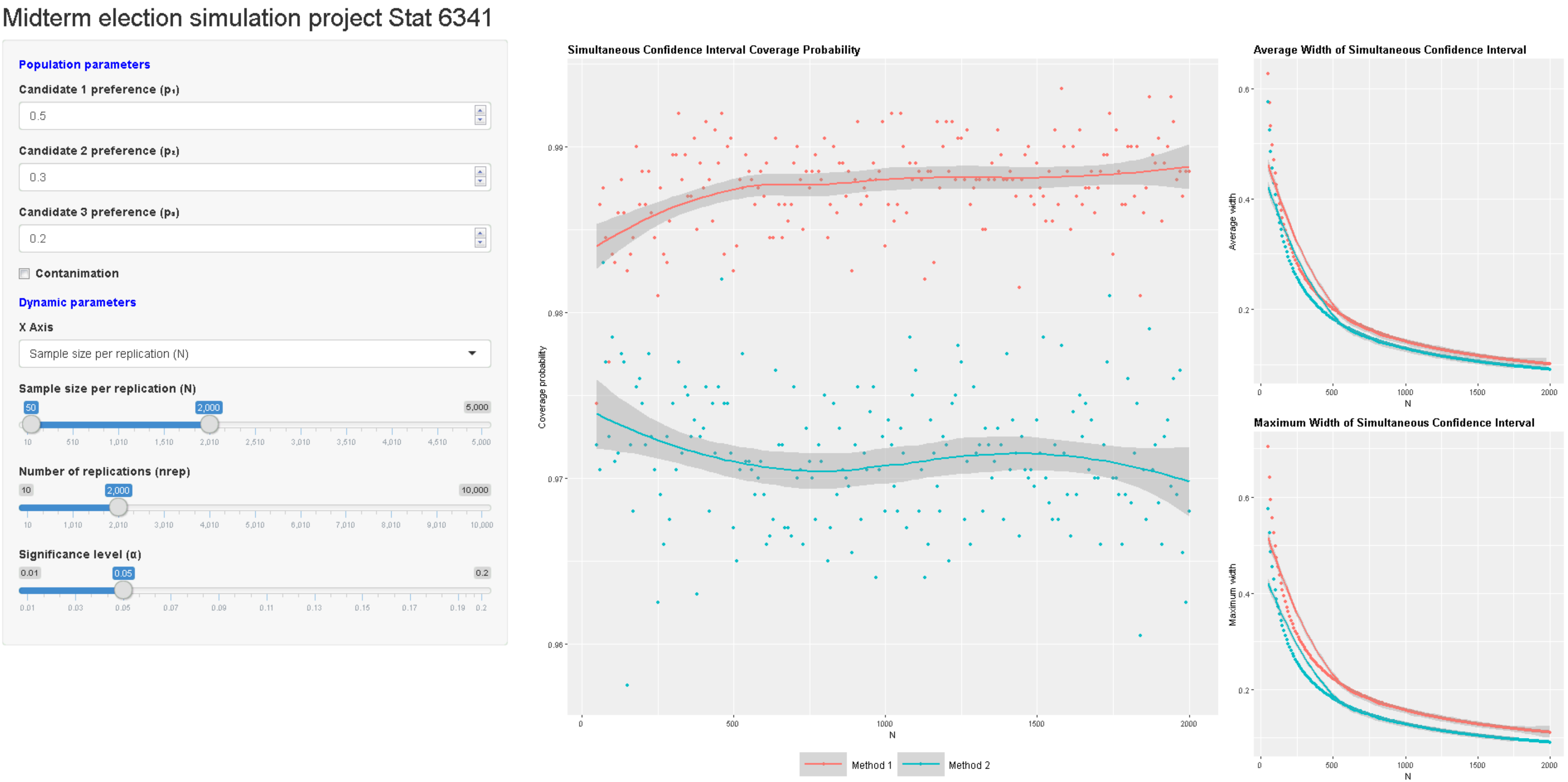
- The estimated maximum simultaneous CI width is $\frac{\sum_t max\_width\_m1_t}{nrep}$ for method 1. Same goes for method 2.

Shiny app:

https://zxynj.shinyapps.io/Midterm_election_simulation_project_Stat_6341/

Shiny app interactive plot

- When there is no contamination, method 1 is better than method 2 but with wider CI.
- CI width decreases as *N* increase.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.5

Candidate 2 preference (p_2): 0.3

Candidate 3 preference (p_3): 0.2

□ Contamination

Dynamic parameters

X Axis: Sample size per replication (N)

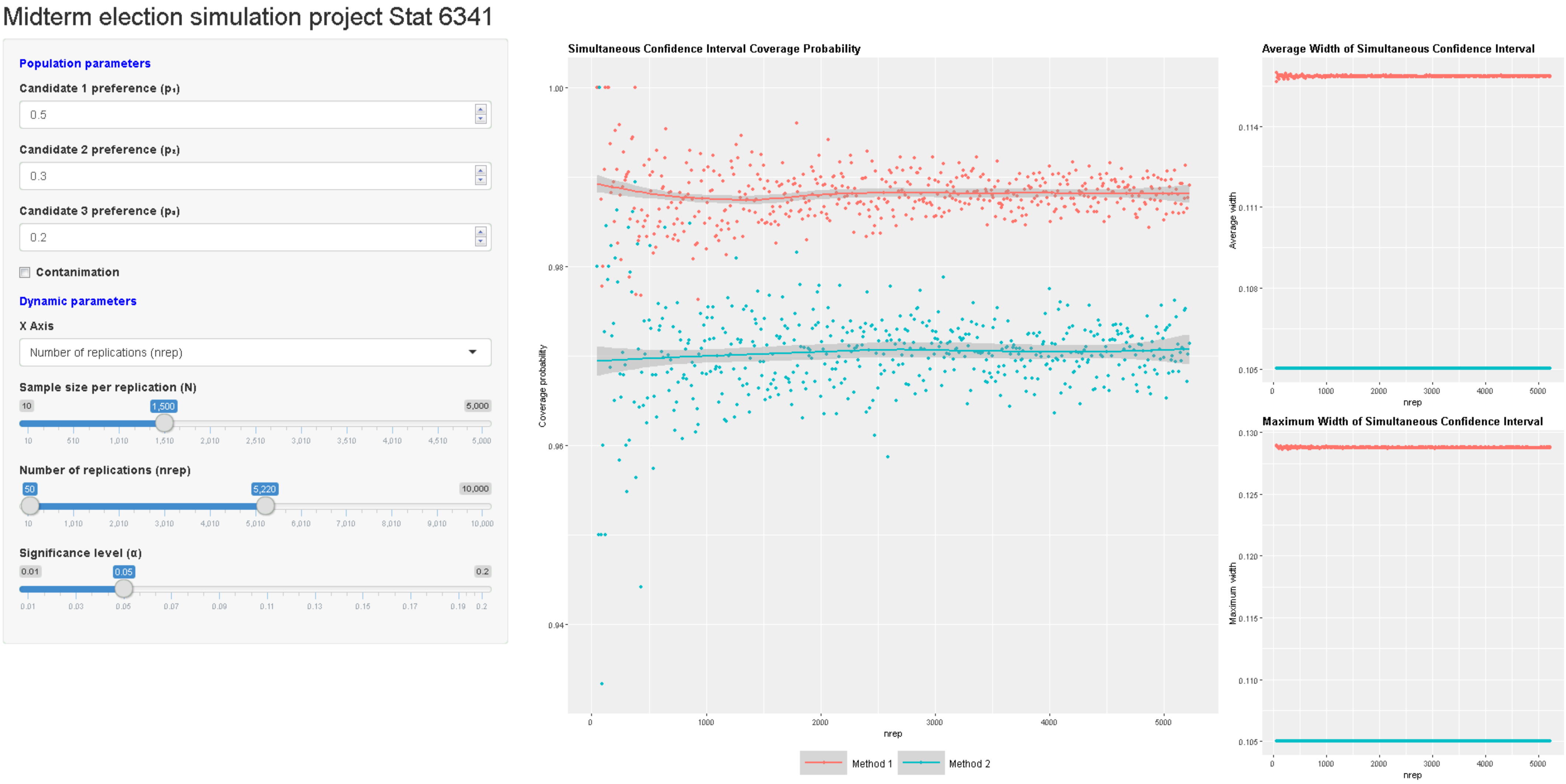
Sample size per replication (N): 50 – 2,000

Number of replications (nrep): 2,000

Significance level (α): 0.05

Shiny app interactive plot

- The variance of the coverage probability decreases as *nrep* increase.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.5

Candidate 2 preference (p_2): 0.3

Candidate 3 preference (p_3): 0.2

□ Contanimation

Dynamic parameters

X Axis: Number of replications (nrep)

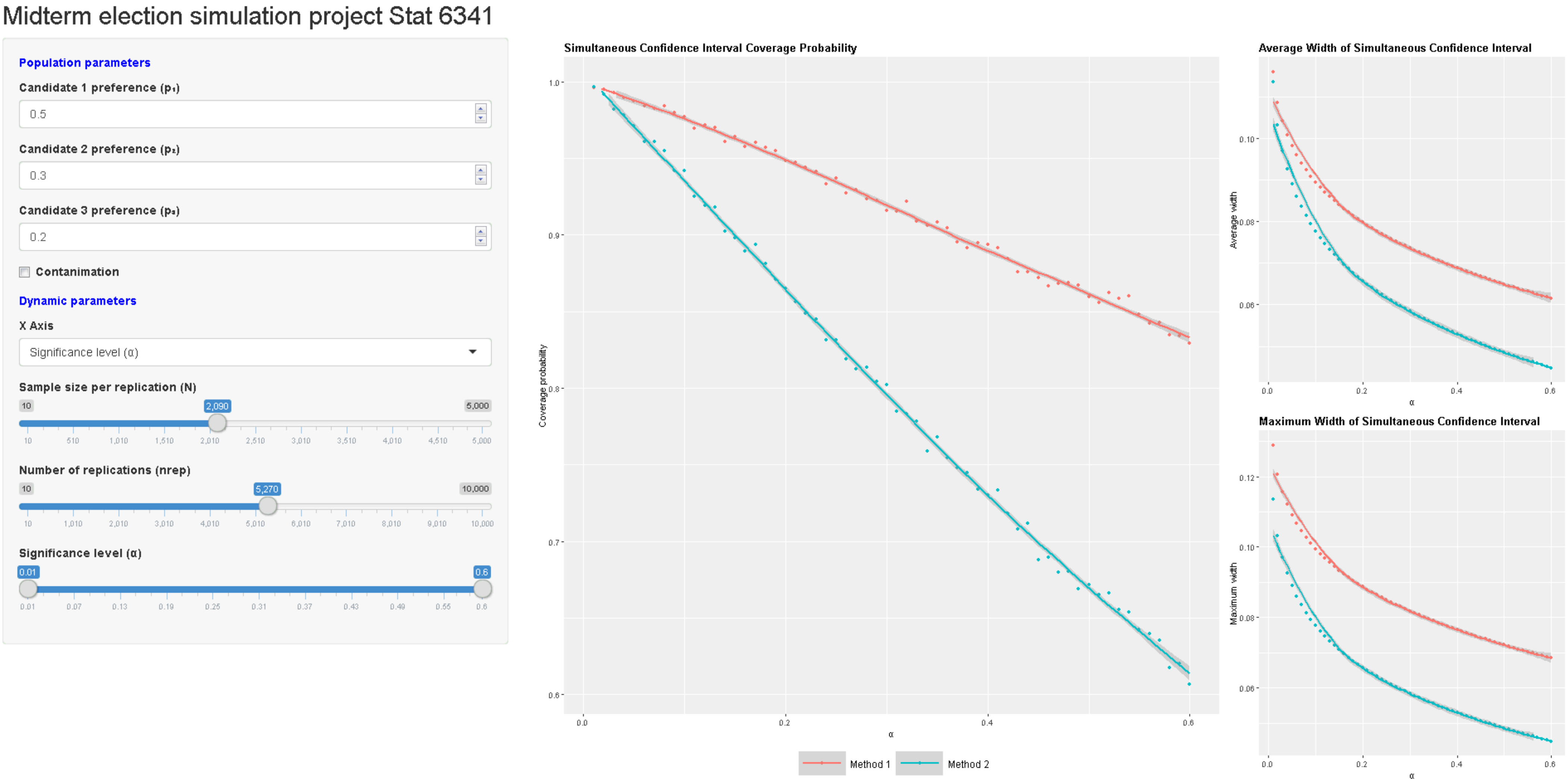
Sample size per replication (N): 1,500 (10 – 5,000)

Number of replications (nrep): 50 – 5,220 (10 – 10,000)

Significance level (α): 0.05 (0.01 – 0.2)

Shiny app interactive plot

- The coverage probability decreases as α increase.
- The coverage probability for Method 2 drops faster than Method 1. This could be explained by Method 2's narrower CI width and faster deacreasing CI width.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.5

Candidate 2 preference (p_2): 0.3

Candidate 3 preference (p_3): 0.2

☐ Contamination

Dynamic parameters

X Axis: Significance level (α)

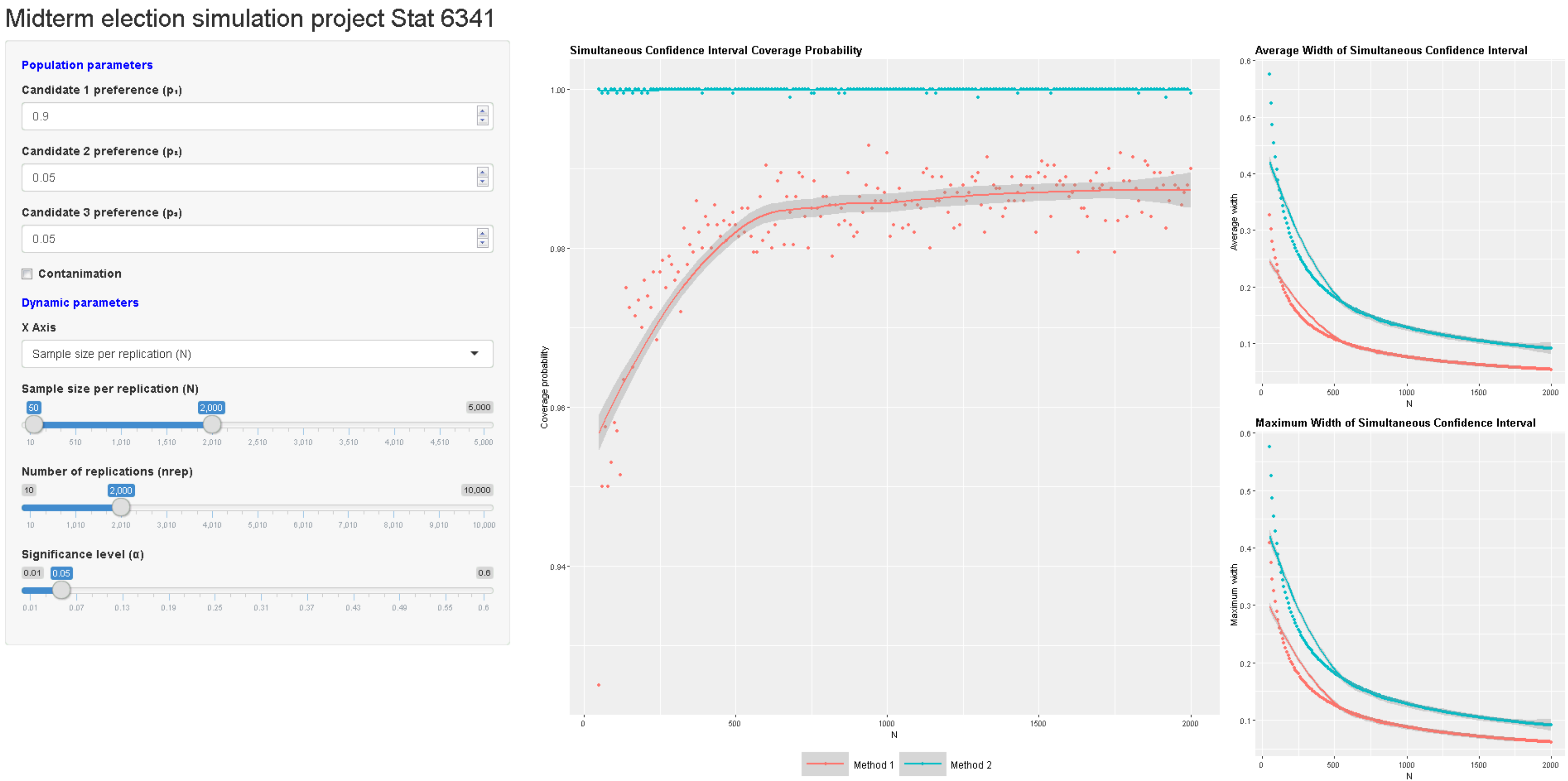
Sample size per replication (N): 2,090

Number of replications (nrep): 5,270

Significance level (α): 0.01 – 0.6

Shiny app interactive plot

- When one candidate dominates the race, the CI width for Method 1 → 0 which cause its coverage probability to perform poorly comparing to Method 1 which has a fixed CI width.
- $d_{i,j} = p_i\_hat + p_j\_hat - (p_i\_hat - p_j\_hat)^2$



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.9

Candidate 2 preference (p_2): 0.05

Candidate 3 preference (p_3): 0.05

☐ Contanimation

Dynamic parameters

X Axis: Sample size per replication (N)

Sample size per replication (N): 50 – 2,000

Number of replications (nrep): 2,000

Significance level (α): 0.05

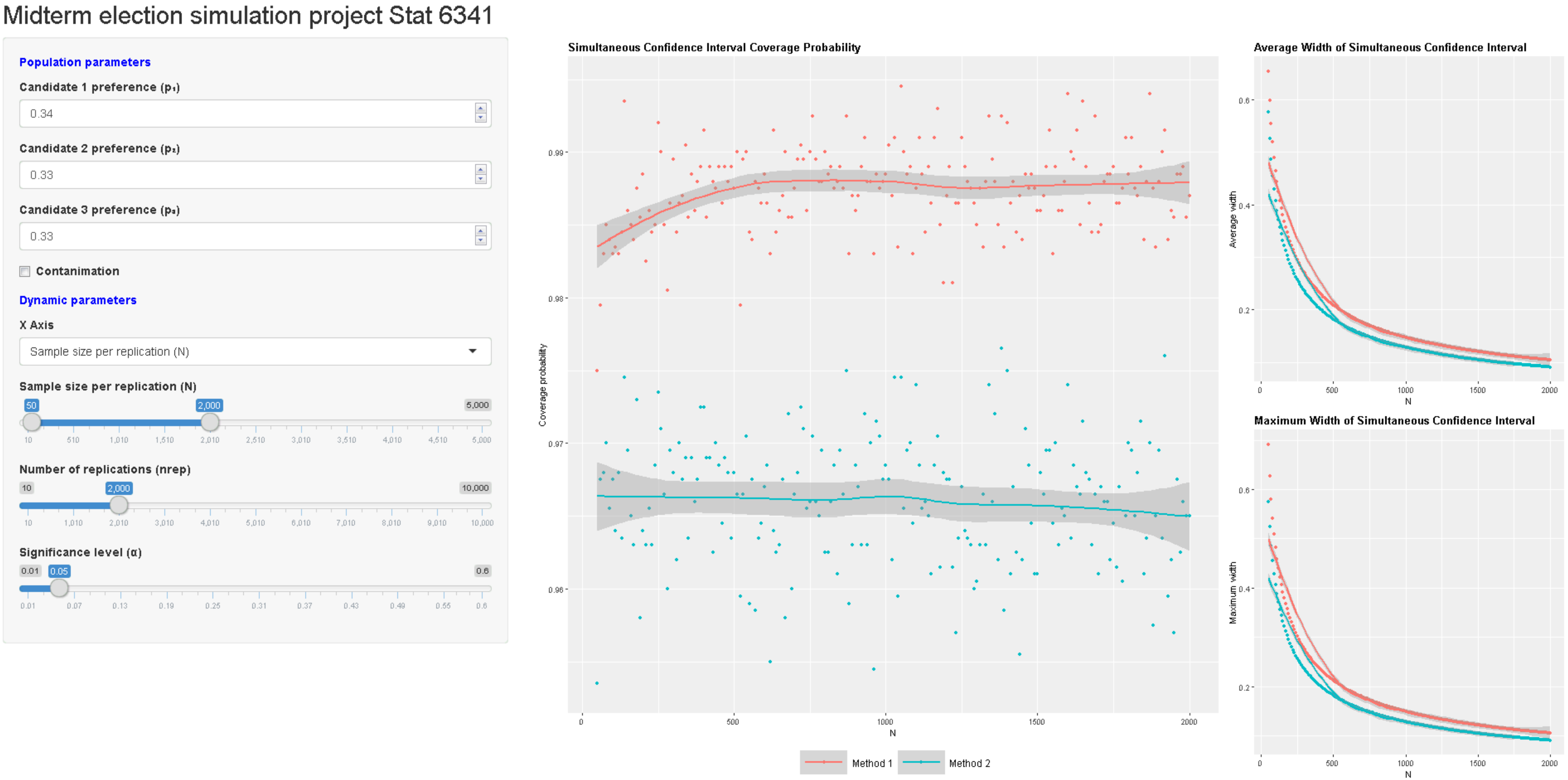
Simultaneous Confidence Interval Coverage Probability

Average Width of Simultaneous Confidence Interval

Maximum Width of Simultaneous Confidence Interval

Shiny app interactive plot

- When all three candidates are very close.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1)

0.34

Candidate 2 preference (p_2)

0.33

Candidate 3 preference (p_3)

0.33

□ Contanimation

Dynamic parameters

X Axis

Sample size per replication (N)

Sample size per replication (N)

50 – 2,000 (range 10 to 5,000)

Number of replications (nrep)

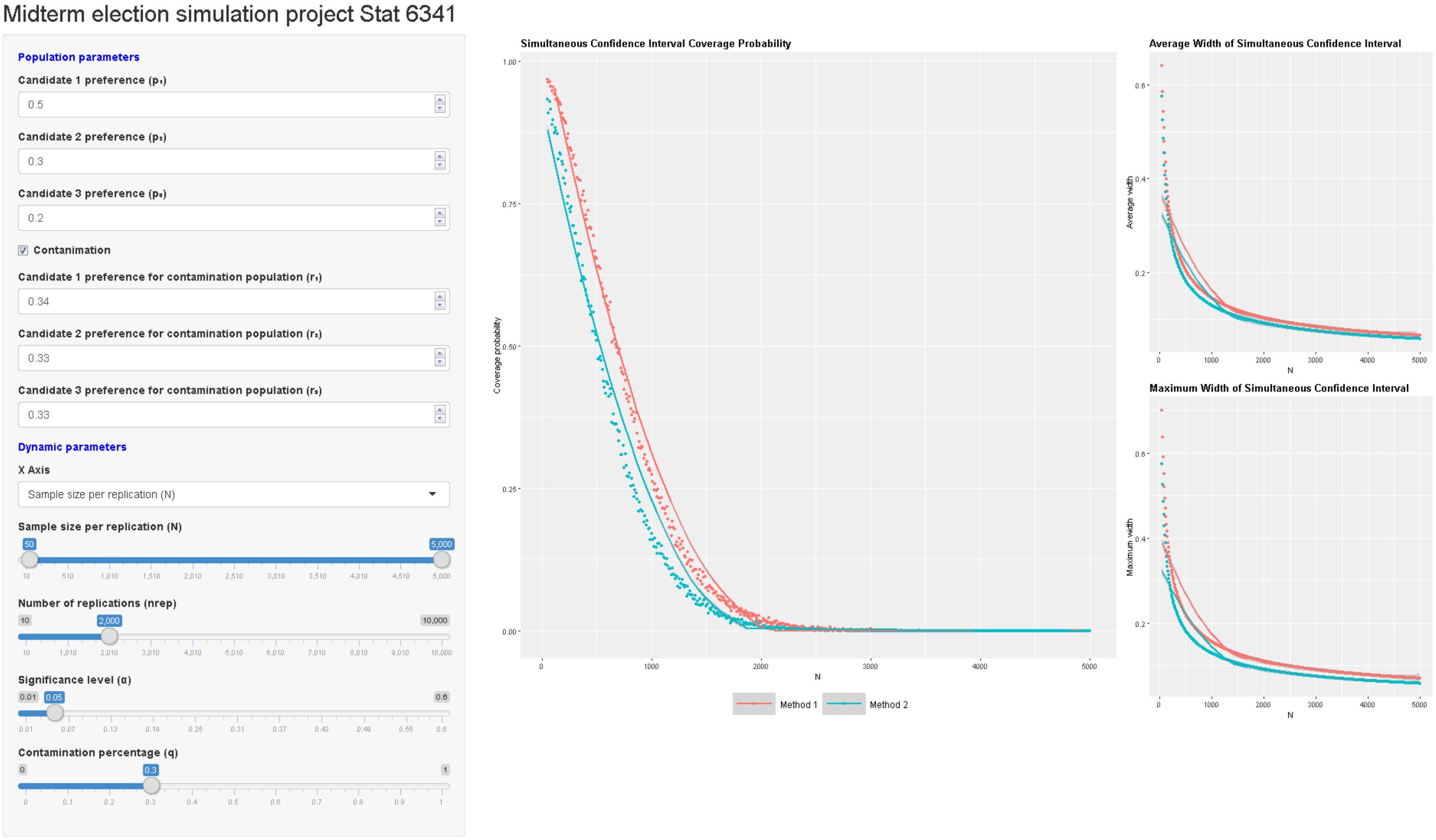
2,000 (range 10 to 10,000)

Significance level (α)

0.05 (range 0.01 to 0.6)

Shiny app interactive plot

- When there is contamination, the CI width was wide enough at the beginning to achieve 90%+ coverage probability.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.5

Candidate 2 preference (p_2): 0.3

Candidate 3 preference (p_3): 0.2

☑ Contanimation

Candidate 1 preference for contamination population (r_1): 0.34

Candidate 2 preference for contamination population (r_2): 0.33

Candidate 3 preference for contamination population (r_3): 0.33

Dynamic parameters

X Axis: Sample size per replication (N)

Sample size per replication (N): 50 – 5,000

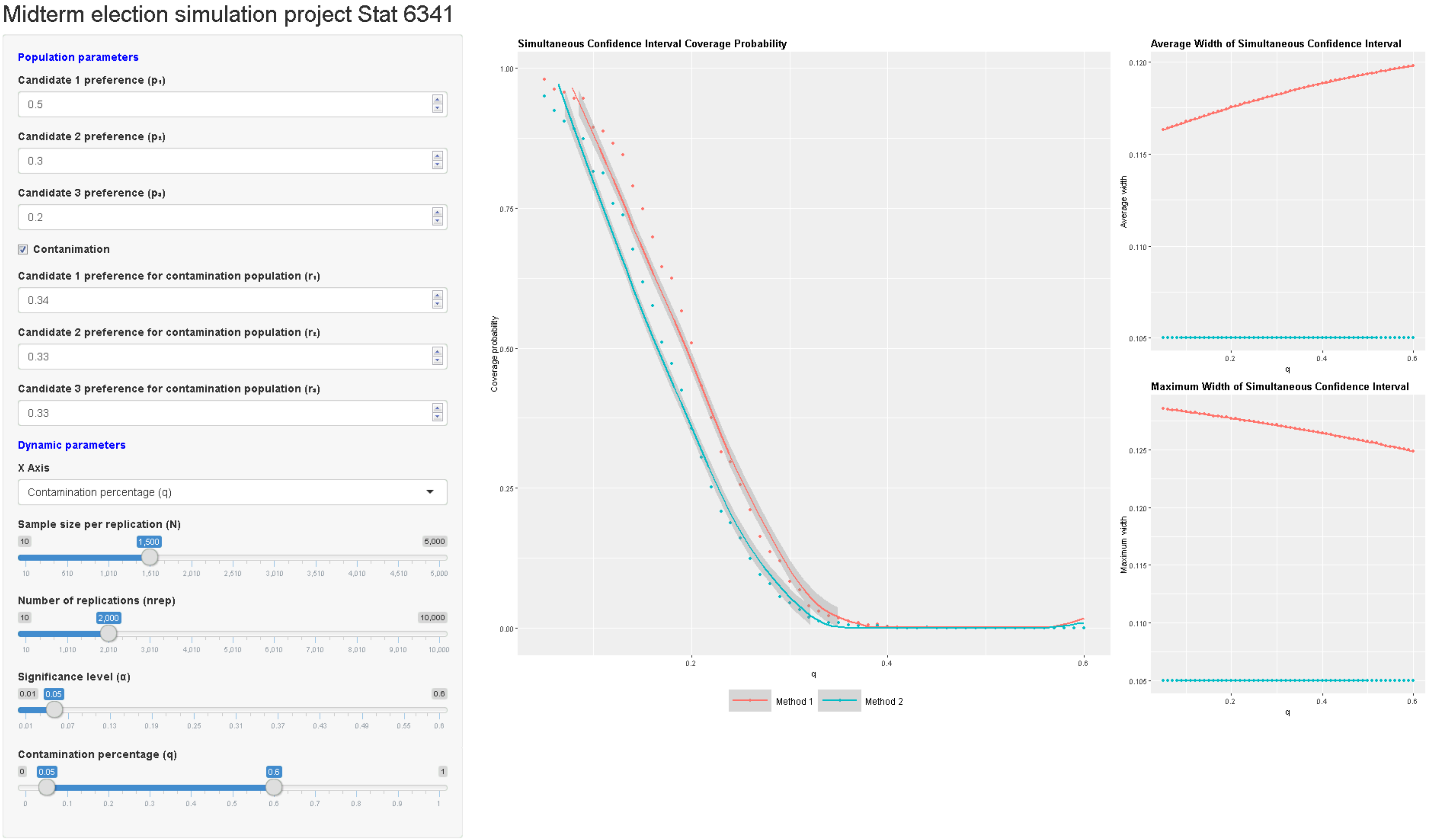
Number of replications (nrep): 2,000

Significance level (α): 0.05

Contamination percentage (q): 0.3

Shiny app interactive plot

- When there is contamination, the coverage probability drops as contamination proportion increases.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1)

0.5

Candidate 2 preference (p_2)

0.3

Candidate 3 preference (p_3)

0.2

☑ Contanimation

Candidate 1 preference for contamination population (r_1)

0.34

Candidate 2 preference for contamination population (r_2)

0.33

Candidate 3 preference for contamination population (r_3)

0.33

Dynamic parameters

X Axis

Contamination percentage (q)

Sample size per replication (N)

10 1,500 5,000

Number of replications (nrep)

10 2,000 10,000

Significance level (α)

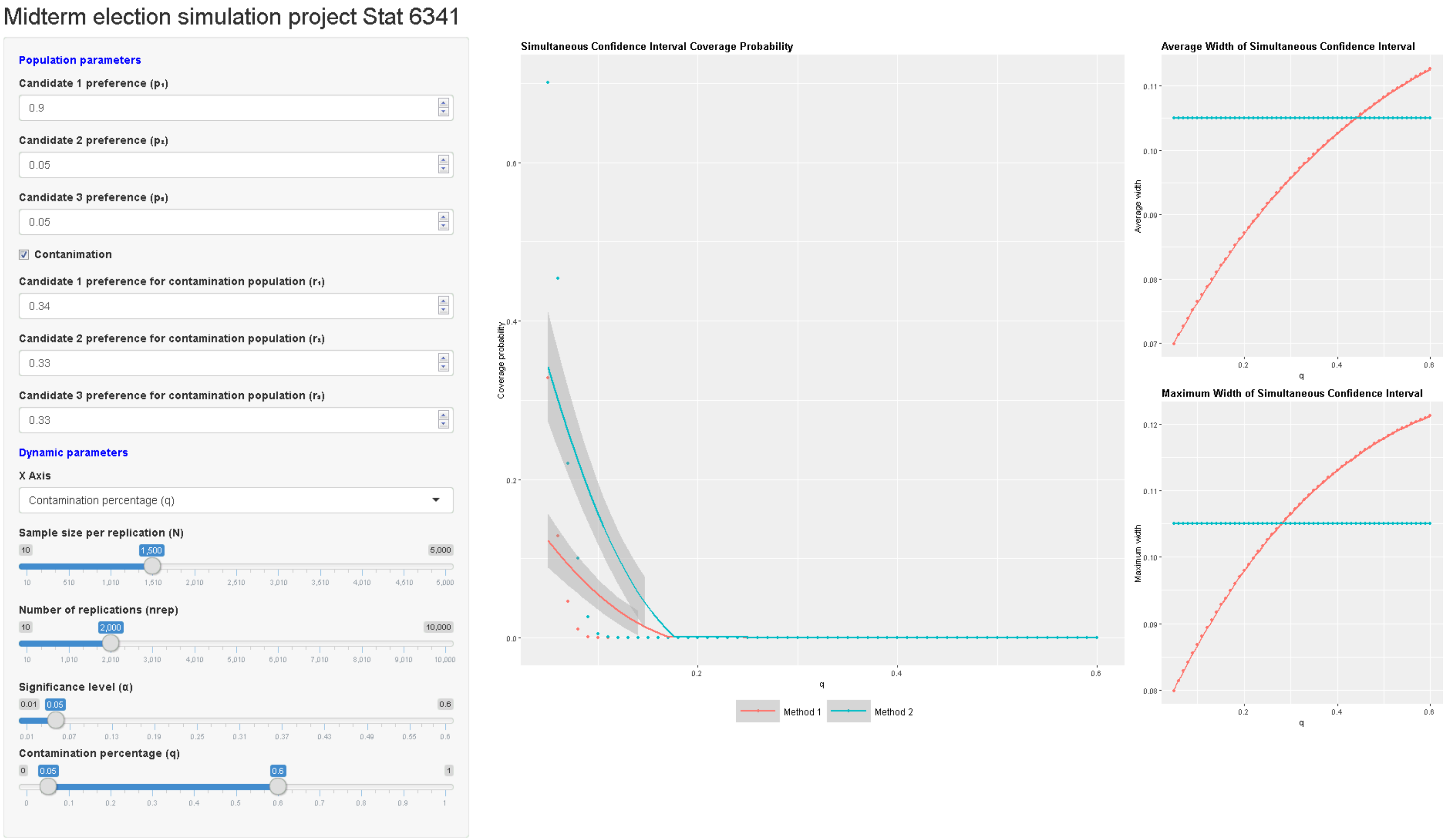
0.01 0.05 0.6

Contamination percentage (q)

0 0.05 0.6 1

Shiny app interactive plot

- When single domination is contaminated with close preference, the CI width for method 1 is smaller than Method 2 as expected.
- As the sample gets more contaminated, the CI center shifts towards the contaminating group's preference probability (close preference) which has a larger CI.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.9

Candidate 2 preference (p_2): 0.05

Candidate 3 preference (p_3): 0.05

☑ Contanimation

Candidate 1 preference for contamination population (r_1): 0.34

Candidate 2 preference for contamination population (r_2): 0.33

Candidate 3 preference for contamination population (r_3): 0.33

Dynamic parameters

X Axis: Contamination percentage (q)

Sample size per replication (N): 1,500 (10 – 5,000)

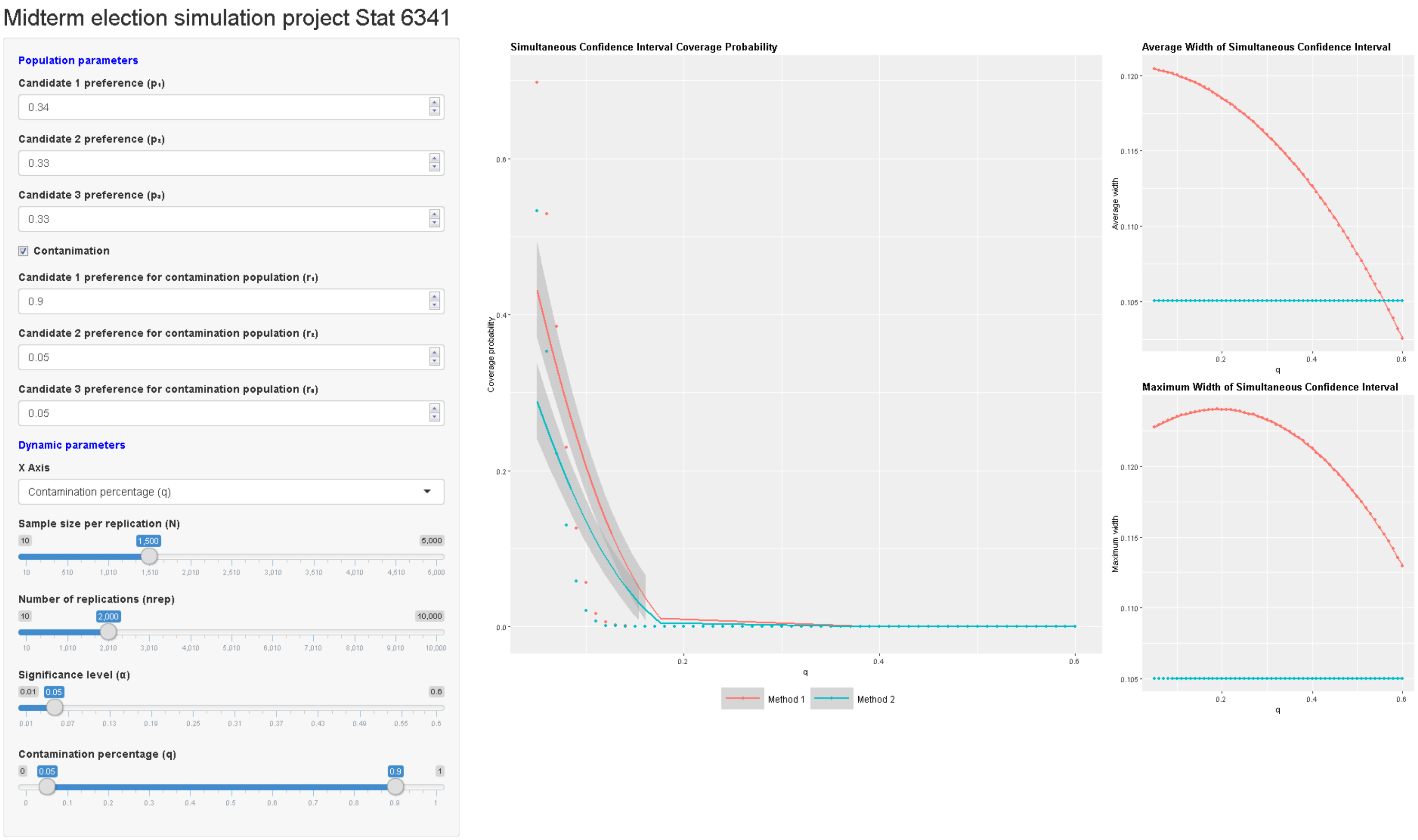
Number of replications (nrep): 2,000 (10 – 10,000)

Significance level (α): 0.05 (0.01 – 0.6)

Contamination percentage (q): 0.05 – 0.6 (0 – 1)

Shiny app interactive plot

- When close preference is contaminated with single domination, as the sample gets more contaminated the CI center shifts towards the contaminating group’s preference probability (single domination) which has a smaller CI.
- Even the overall trend of maximum CI with is decreasing, it actually goes up a little bit as the mixed p_1, p_2, p_3 becomes a little bit different.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1): 0.34

Candidate 2 preference (p_2): 0.33

Candidate 3 preference (p_3): 0.33

☑ Contanimation

Candidate 1 preference for contamination population (r_1): 0.9

Candidate 2 preference for contamination population (r_2): 0.05

Candidate 3 preference for contamination population (r_3): 0.05

Dynamic parameters

X Axis: Contamination percentage (q)

Sample size per replication (N): 1,500 (10 – 5,000)

Number of replications (nrep): 2,000 (10 – 10,000)

Significance level (α): 0.05 (0.01 – 0.6)

Contamination percentage (q): 0.05 – 0.9 (0 – 1)

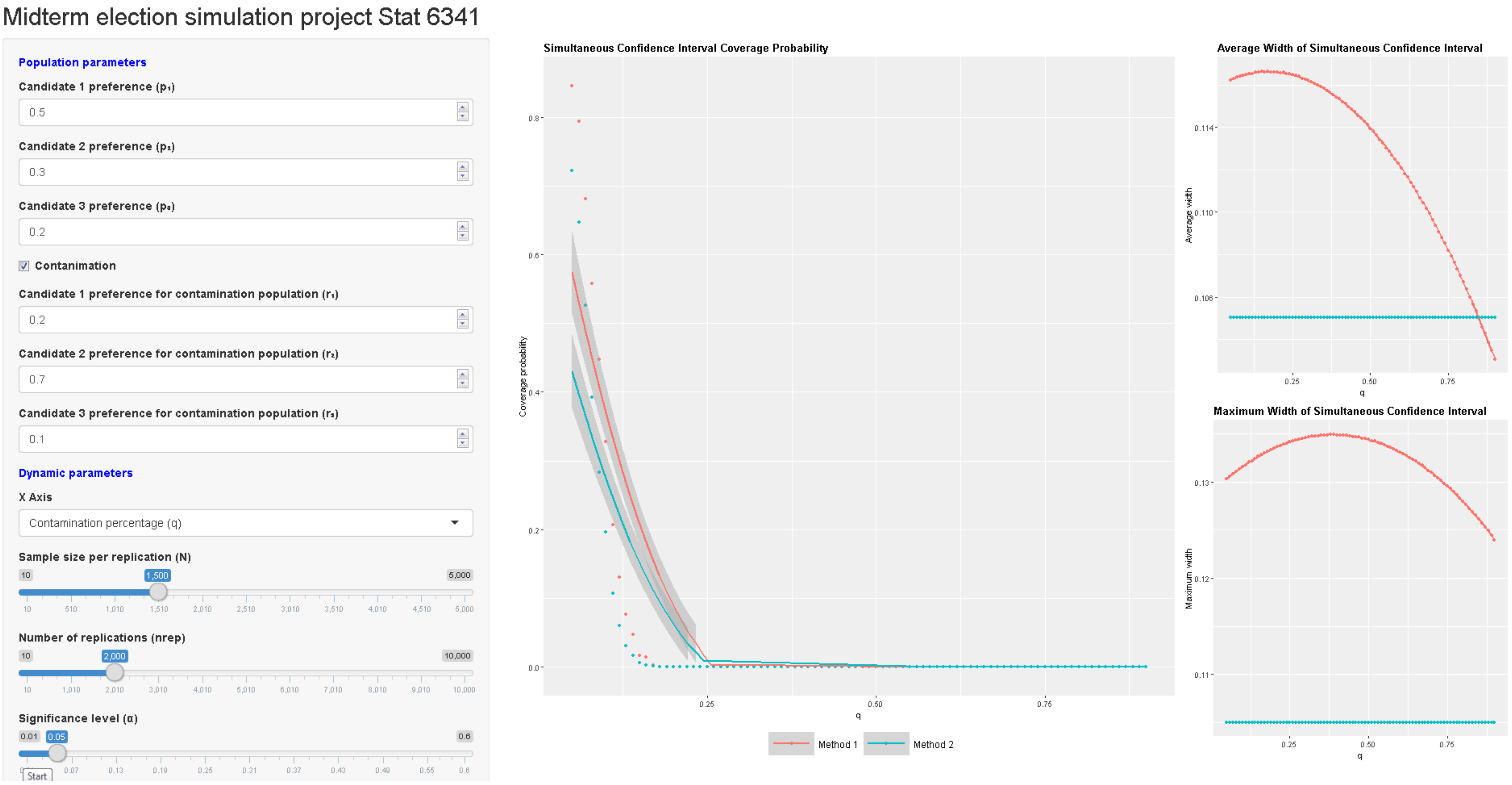
Simultaneous Confidence Interval Coverage Probability

Average Width of Simultaneous Confidence Interval

Maximum Width of Simultaneous Confidence Interval

Shiny app interactive plot

- Compare p_1, p_2, p_3 mixed with r_1, r_2, r_3 and p_1, p_2, p_3 mixed with r_1, r_3, r_2.
- Coverage probability becomes 0 at q = 0.25 and average CI width has a small ascent right after mixing.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p_1)

0.5

Candidate 2 preference (p_2)

0.3

Candidate 3 preference (p_3)

0.2

☑ Contamination

Candidate 1 preference for contamination population (r_1)

0.2

Candidate 2 preference for contamination population (r_2)

0.7

Candidate 3 preference for contamination population (r_3)

0.1

Dynamic parameters

X Axis

Contamination percentage (q)

Sample size per replication (N)

10 1,500 5,000

Number of replications (nrep)

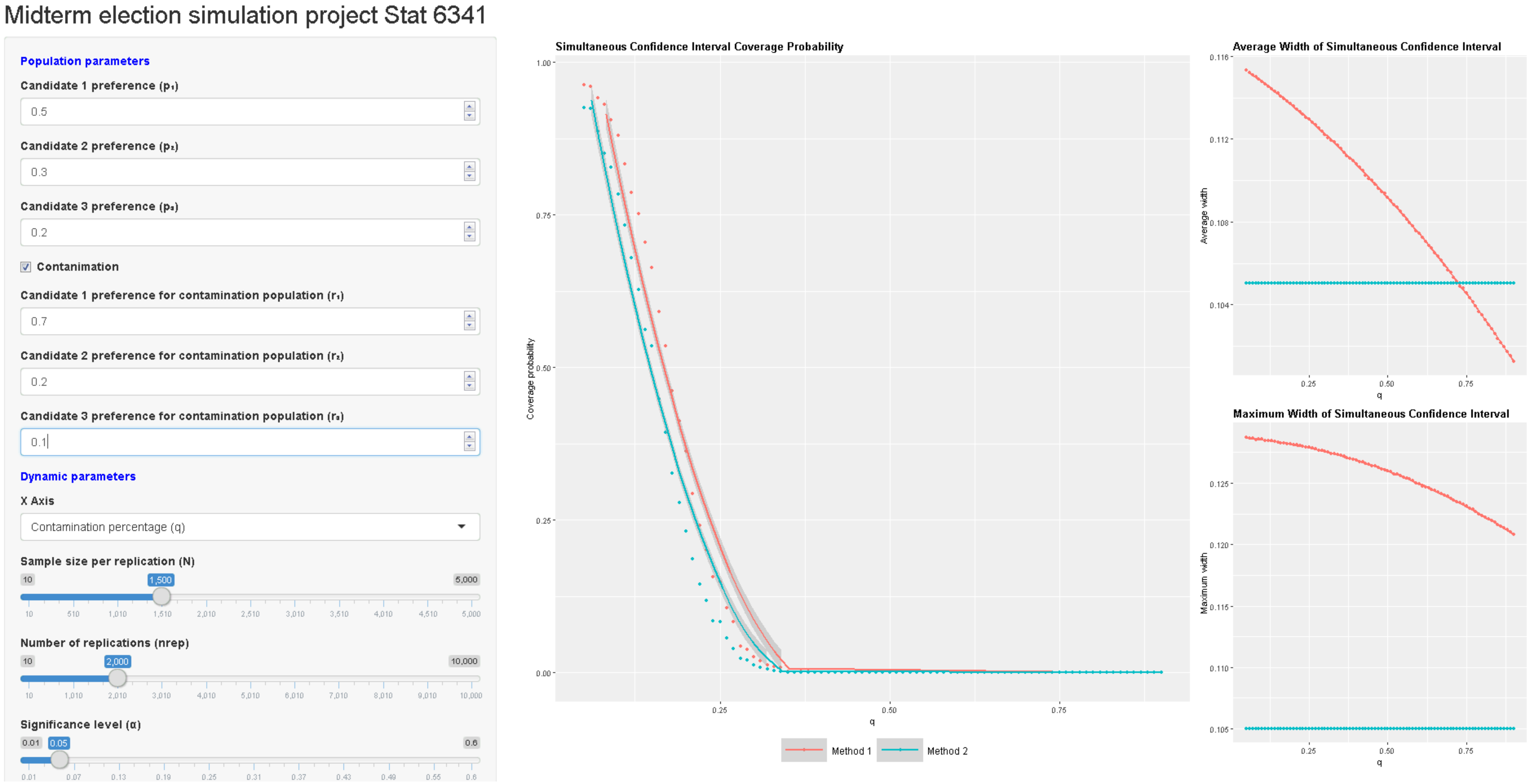
10 2,000 10,000

Significance level (α)

0.01 0.05 0.6

Shiny app interactive plot

- Compare p_1, p_2, p_3 mixed with r_1, r_2, r_3 and p_1, p_2, p_3 mixed with r_1, r_3, r_2.
- Coverage probability becomes 0 at q =0.35.



Midterm election simulation project Stat 6341

Population parameters

Candidate 1 preference (p$_1$): 0.5

Candidate 2 preference (p$_2$): 0.3

Candidate 3 preference (p$_3$): 0.2

☑ Contamination

Candidate 1 preference for contamination population (r$_1$): 0.7

Candidate 2 preference for contamination population (r$_2$): 0.2

Candidate 3 preference for contamination population (r$_3$): 0.1

Dynamic parameters

X Axis: Contamination percentage (q)

Sample size per replication (N): 1,500

Number of replications (nrep): 2,000

Significance level (α): 0.05