Numerical Linear Algebra and Statistical Computing

Mathematical Sciences

Project 1

Author:
Qinyi Zhou
Akash Roy
Dipnil Chakraborty
Randy Suarez Rodes

Supervisor:
Dr. Larry Ammann

November 6, 2018

1 Introduction

During each election year we are inundated with the results of political preference polls. The media presents the results of these polls with little or no discussion of their accuracy. TV news reports typically present a set of poll percentages as if they were the actual population percentages, while newspapers may include a small-font footnote stating that the results have an error of $\pm(3-5)\%$. Such statements are based on the conservative bound on the s.d. of a sample proportion and ignore several issues that can impact the results significantly. This project is intended to examine some of those issues.

Note that the percentages for each candidate are not independent, and so confidence intervals for each percentage would not be independent. Suppose that the poll asks respondents how likely they are to vote and then labels each respondent as either likely to vote or not likely to vote. The final poll report is based only on those who are labeled as likely to vote.

There are two types of populations to be examined. First scenario is the preferences of those who claim they will vote but don't vote and those who actually vote are the same. Other one is the preferences of those who claim they will vote but don't vote and those who actually vote are not the same.

2 Methods

Three candidates are running for a political office and a poll is conducted which randomly selects a sample of size n from the population of eligible voters and then obtains the sample percentages for each candidate. For each type of population, simulate sample sizes of N = 500, 1000, 1500 for the polls and use 2000 replications for the simulations. In the first part of this project we examined two sets of true population proportions: (0.5, 0.3, 0.2) and (0.5, 0.48, 0.02). Then for the second part, only one true set of true proportions are considered: (0.5, 0.25, 0.25). The samples for this part are then mixing randomly with non-voters that have a set of true proportions: (0.1, 0.45, 0.45). A binomial random variable using rbinom was used to determine the mixing rate (number of non-voters in each sample). Results will be shown across different mixing rates of: 0.1, 0.2, 0.3, 0.4, and 0.5.

The two methods for calculating 95% confidence intervals are shown below:

2.1 Method GB

A set of approximate $1-\alpha$ confidence intervals for $\Delta_{i,j}$, $1 \leq i < j \leq m$, is

$$\delta_{i,j} \pm \sqrt{\frac{Ad_{i,j}}{N}},$$

where

$$\begin{aligned} A &= \chi^2_{M-1}(\alpha/M), \\ d_{i,j} &= \hat{p}_i + \hat{p}_j - (\delta_{i,j})^2 \end{aligned}$$

and $\chi^2_n(\alpha)$ denotes the $1-\alpha$ quantile from the chi-square distribution with n d.f. This critical value can be obtained in R by cv = qchisq(1-alpha,n)

2.2 Method FS

The confidence intervals are given by

$$\delta_{i,j} \pm \frac{a}{\sqrt{N}},$$

where a is the solution to the equation:

$$1 - 2[1 - \Phi(a)] - 4[m-2][1 - \Phi(a\sqrt{2})] = 1 - \alpha,$$

3 Part 1

The following two situations are examined.

- (a) (p_1, p_2, p_3) = (0.5, 0.3, 0.2)
- (b) (p_1,p_2, p_3) = (0.5, 0.48, 0.02)

Table 1: Simultaneous coverage probability for Part 1

Proportions	**Sample size**	**Method GB**	**Method FS**
(0.5,0.3,0.2)	500	0.9860	0.9655
(0.5,0.3,0.2)	1000	0.9860	0.9690
(0.5,0.3,0.2)	1500	0.9885	0.9700
(0.5,0.48,0.02)	500	0.9865	0.9565
(0.5,0.48,0.02)	1000	0.9920	0.9620
(0.5,0.48,0.02)	1500	0.9905	0.9640

Conclusion from Part 1:

- The simultaneous coverage probabilities obtained using method GB are higher than the simultaneous coverage probabilities obtained using method FS for both situations.

- For the first proportion, when the sample sizes increases the simultaneous coverage probabilities increases for both methods.

- But there are no obvious pattern for the other proportion when the sample sizes increases.

4 Part 2

The following two subpopulations were used.

- (a) (p_1, p_2, p_3) = (0.5, 0.25, 0.25), the population proportions those of those who actually vote.

- (b) (r_1,r_2, r_3) = (0.1, 0.45, 0.45), the population proportions of those who claim they will vote when contacted, but don't actually vote during the election.

Let q be the proportion of those labeled as likely to vote who do not vote.

Table 2: Simultaneous Coverage probability for Part 2

Sample Size	Method	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
500	GB	0.8445	0.2215	0.0055	0.0000	0.0000
	FS	0.7060	0.1165	0.0020	0.0000	0.0000
1000	GB	0.6075	0.0160	0.0000	0.0000	0.0000
	FS	0.4215	0.0040	0.0000	0.0000	0.0000
1500	GB	0.3950	0.0010	0.0000	0.0000	0.0000
	FS	0.2260	0.0000	0.0000	0.0000	0.0000

Conclusion from Part 2:

- For each q and n, method GB gives higher simultaneous coverage probabilities than method FS.
- When q increases, simultaneous coverage probabilities decrease for both the methods.
- When n increases, simultaneous coverage probabilities decrease for both the methods.

5 Improvements

Table 3: **Coverage probability for Part 1 with Improvement**

		N=500	N=1000	N=1500
Item a)	GB	0.9860	0.9860	0.9885
	FS	0.9655	0.9690	0.9700
	GB.logratio	0.9890	0.9880	0.9895

		N=500	N=1000	N=1500
Item b)	GB	0.9865	0.9920	0.9905
	FS	0.9565	0.9620	0.9640
	GB.logratio	0.9870	0.9910	0.9880

Table 4: **Coverage probability for Part 2 with Improvement**

Sample Size	Method	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
500	GB	0.8445	0.2215	0.0055	0.0000	0.0000
	FS	0.7060	0.1165	0.0020	0.0000	0.0000
	GB.logratio	0.8495	0.2455	0.0080	0.0000	0.0000
1000	GB	0.6075	0.0160	0.0000	0.0000	0.0000
	FS	0.4215	0.0040	0.0000	0.0000	0.0000
	GB.logratio	0.6380	0.0220	0.0000	0.0000	0.0000
1500	GB	0.3950	0.0010	0.0000	0.0000	0.0000
	FS	0.2260	0.0000	0.0000	0.0000	0.0000
	GB.logratio	0.4285	0.0015	0.0000	0.0000	0.0000

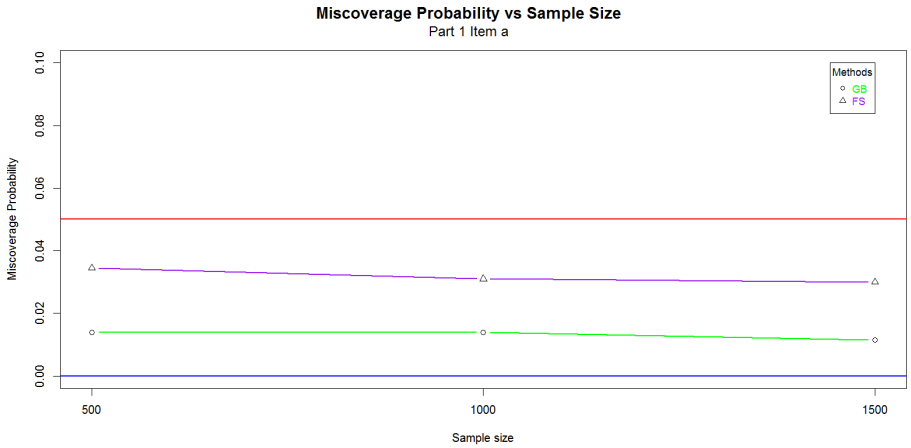

Figure 1: $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$

6 Conclusion

- Polls need to be careful about how they include or exclude people based on their response to the question "are you voting or not?".
- They should ask more questions so as to have a better prediction. According to the results it is clear that contamination affects the results of the polls.
- If we compute individual coverage probabilities then only it is possible to explain whether the poll is favourable to any candidate; otherwise not.

7 Graphical Analysis

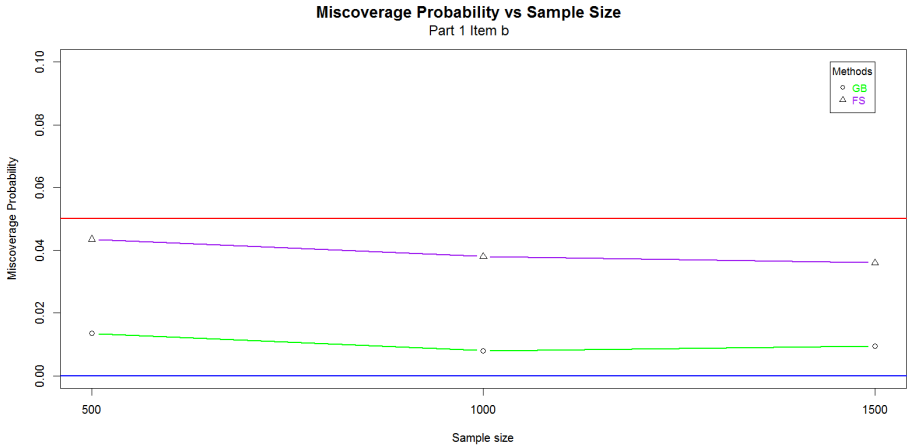

Figure 2: $(p_1, p_2, p_3) = (0.5, 0.48, 0.02)$

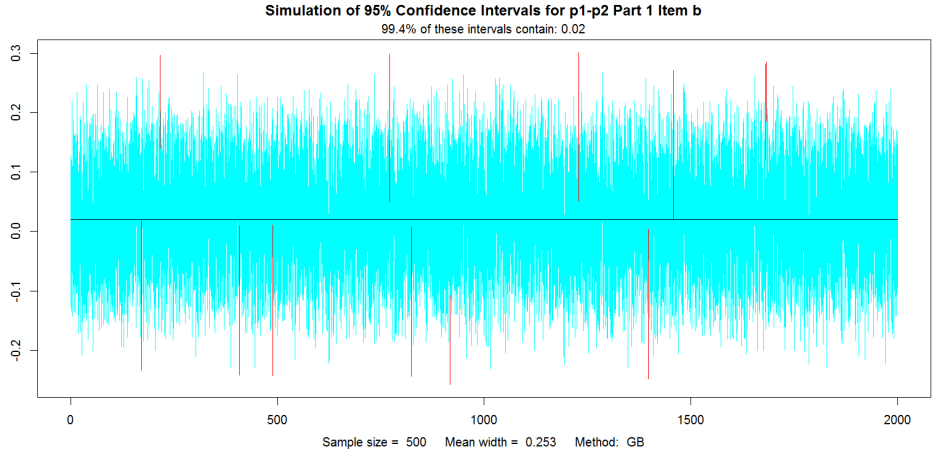

Figure 3: GB Confidence interval for p_1-p_2 for N=500(PART 1(b))

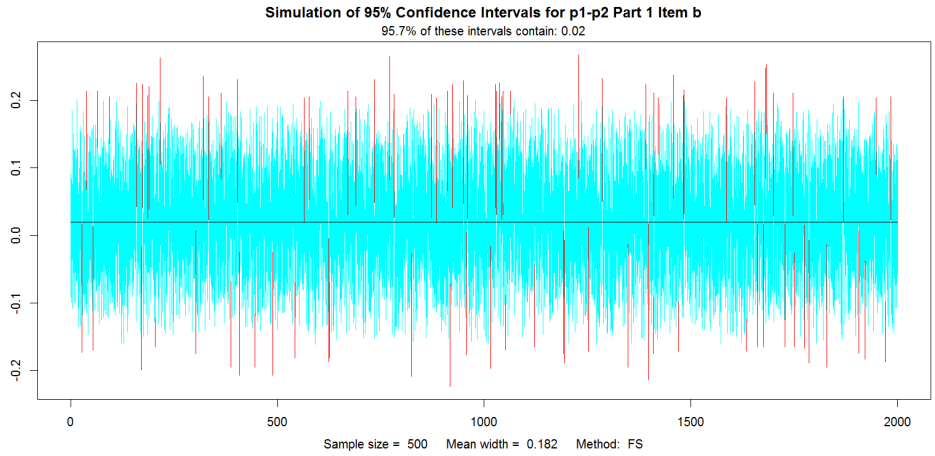

Figure 4: FS Confidence interval for p_1-p_2 for N=500(PART 1(b))

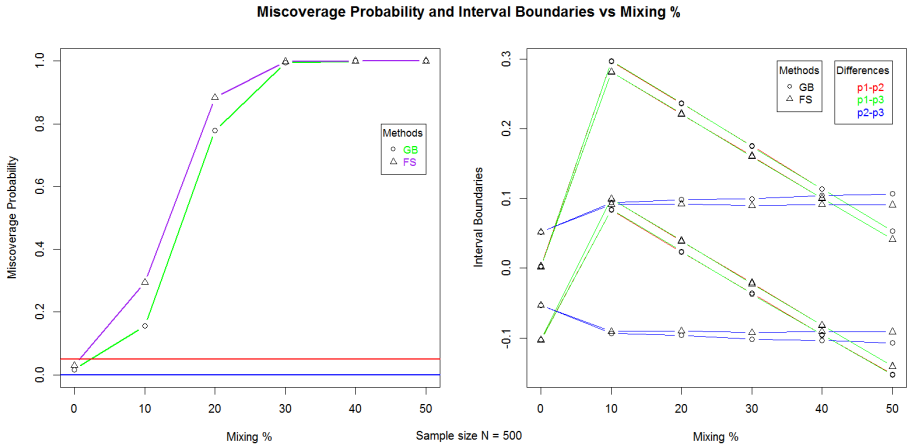

Figure 5: Miscoverage Probability vs Mixing rate for N=500

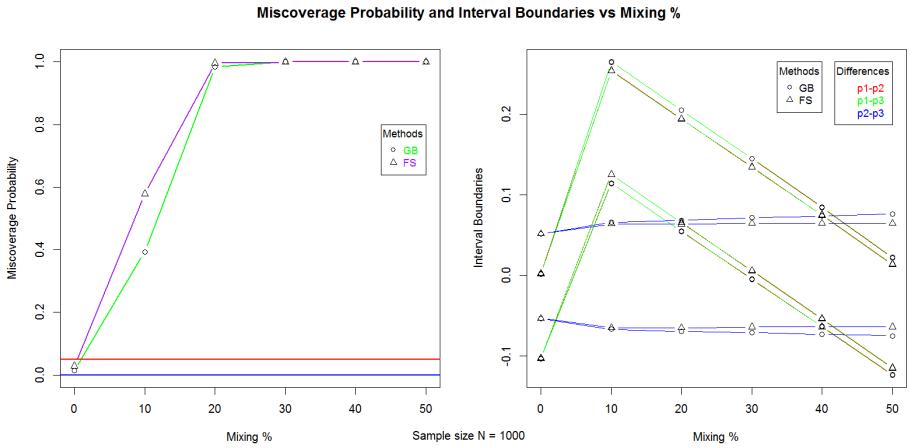

Figure 6: Miscoverage Probability vs Mixing rate for N=1000

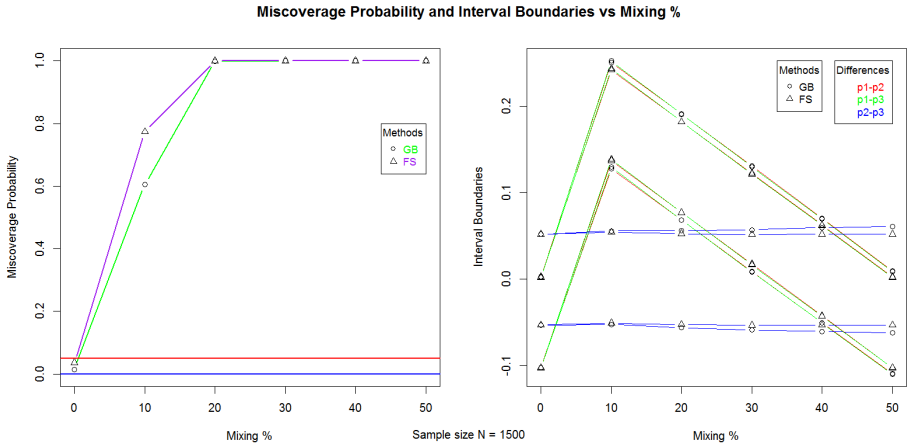

Figure 7: Miscoverage Probability vs Mixing rate for N=1500

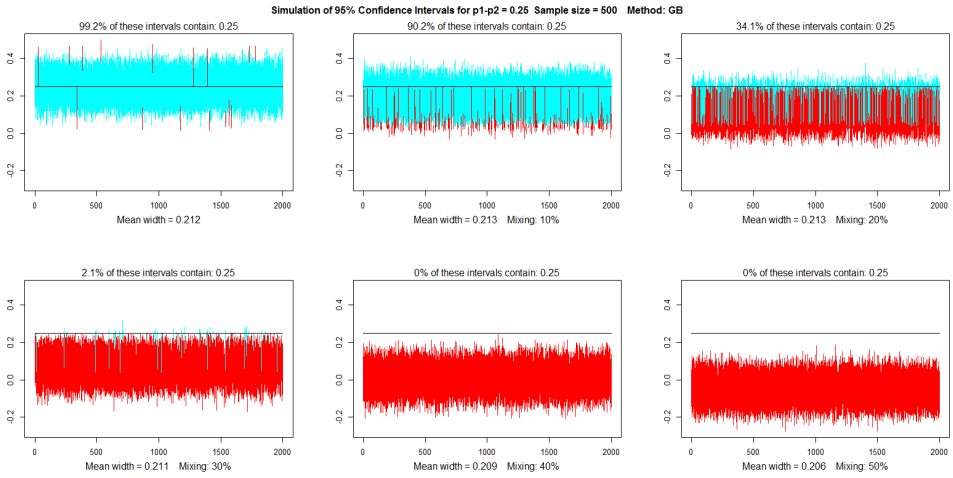

Figure 8: GB Confidence interval for p_1-p_2 for N=500(PART 2)

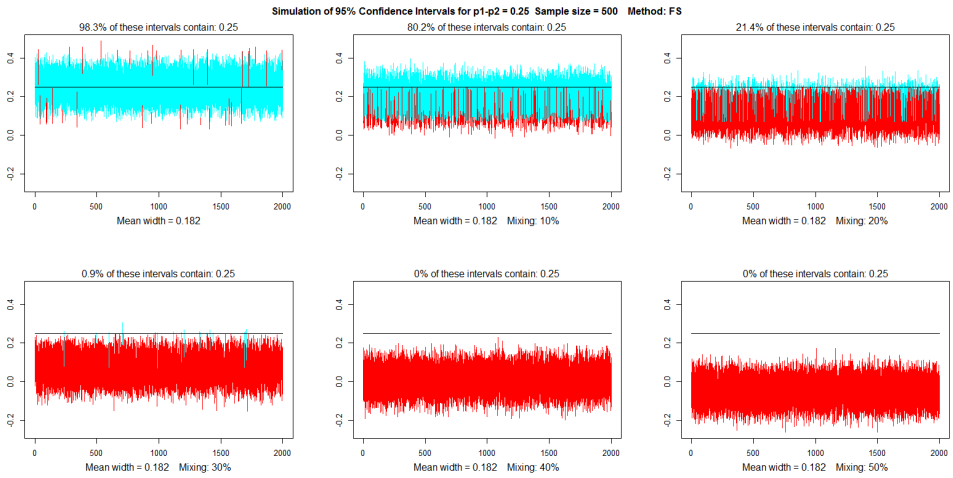

Figure 9: FS Confidence interval for p_1-p_2 for N=500(PART 2)

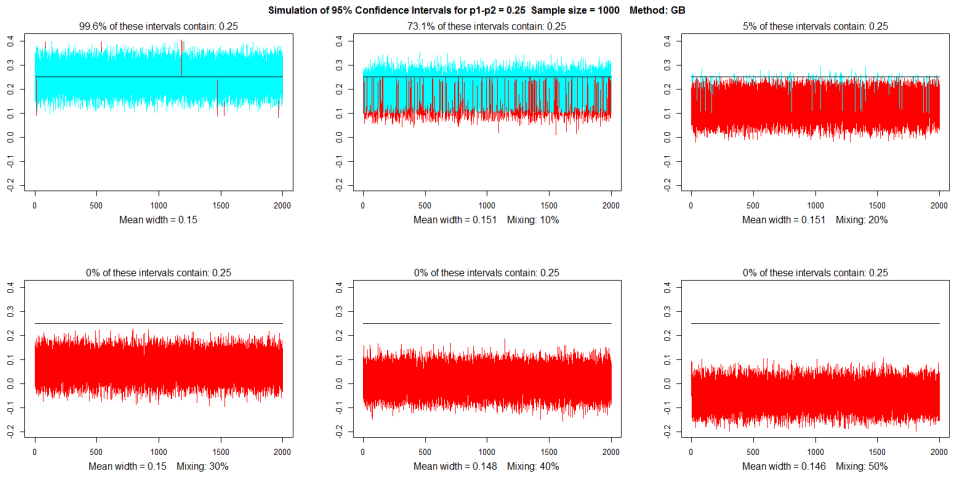

Figure 10: GB Confidence interval for p_1-p_2 for N=1000(PART 2)

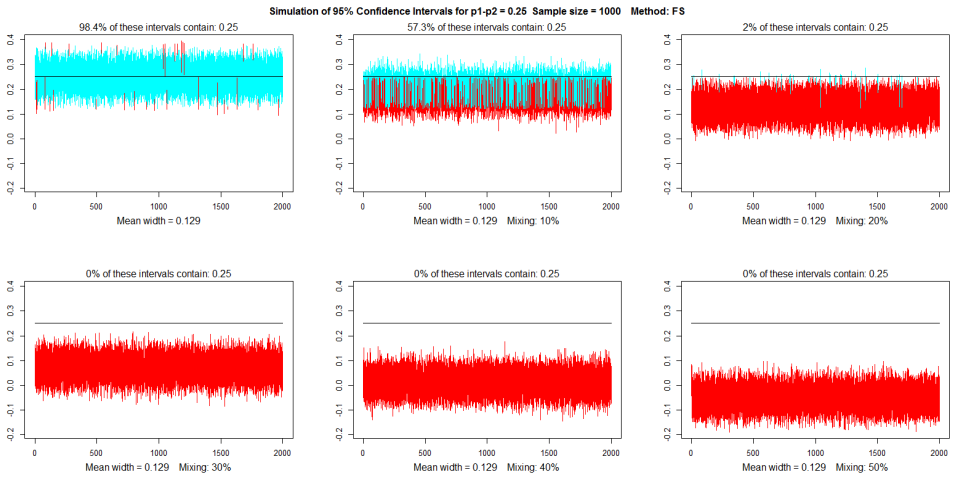

Figure 11: FS Confidence interval for p_1-p_2 for N=1000(PART 2)

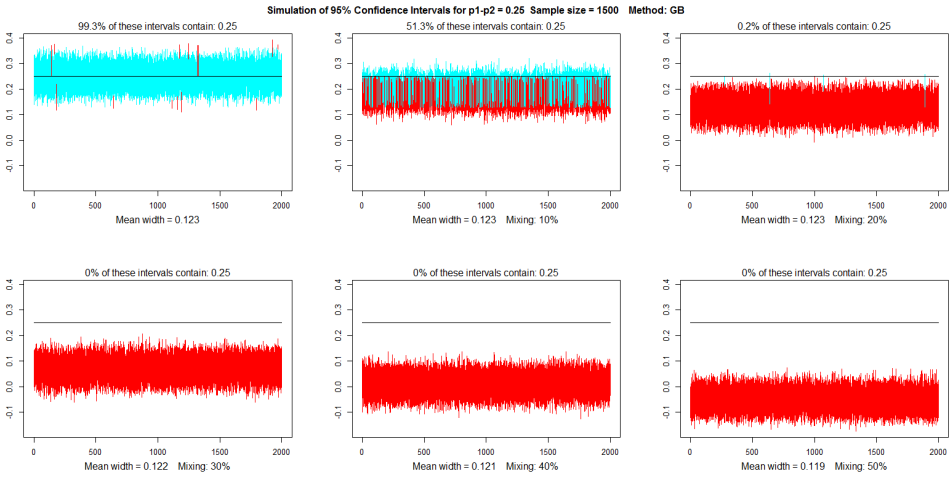

Figure 12: GB Confidence interval for p_1-p_2 for N=1500(PART 2)

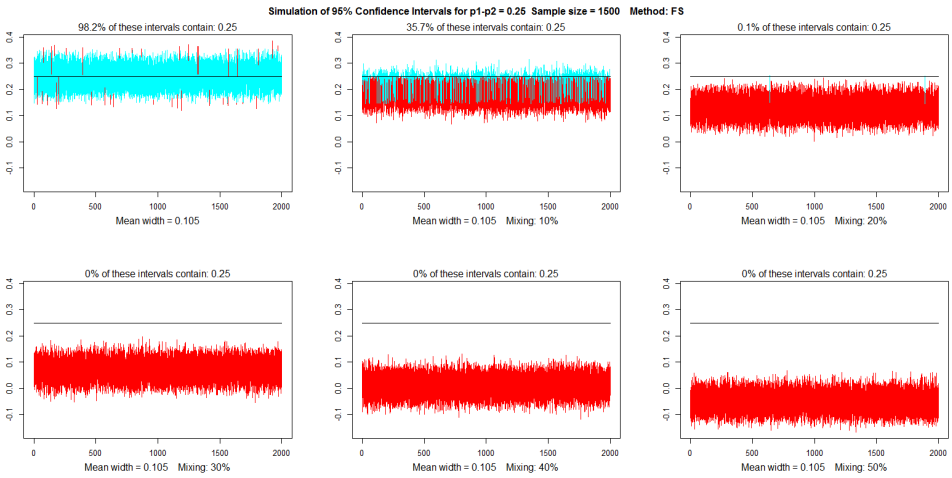

Figure 13: FS Confidence interval for p_1-p_2 for N=1500(PART 2)

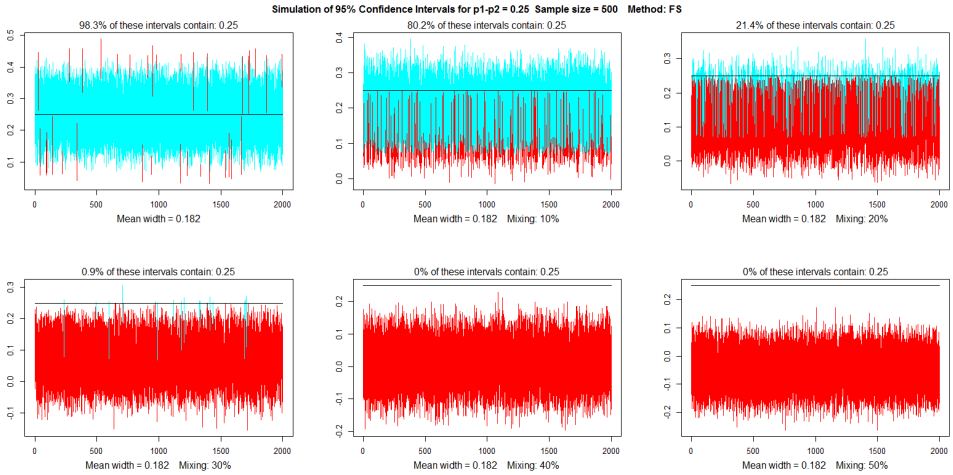

Figure 14: FS CIs for p_1-p_2 for N=500(PART 2) with modified scale

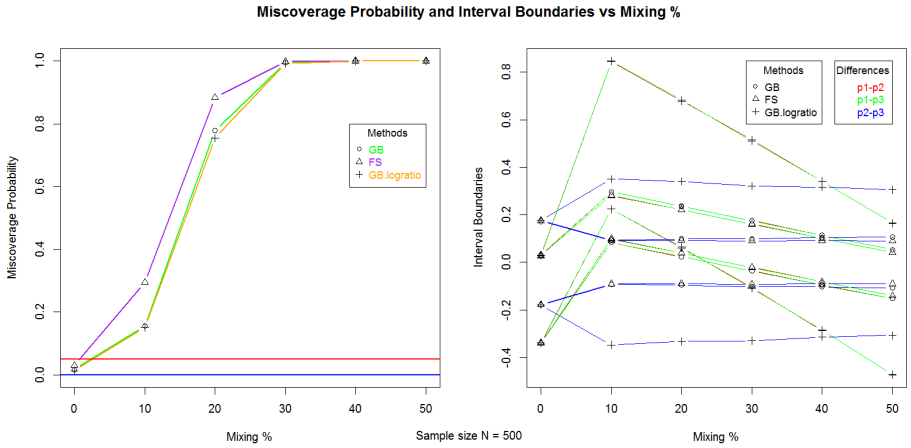

Figure 15: Miscoverage Probability vs Mixing rate with improvement for N=500

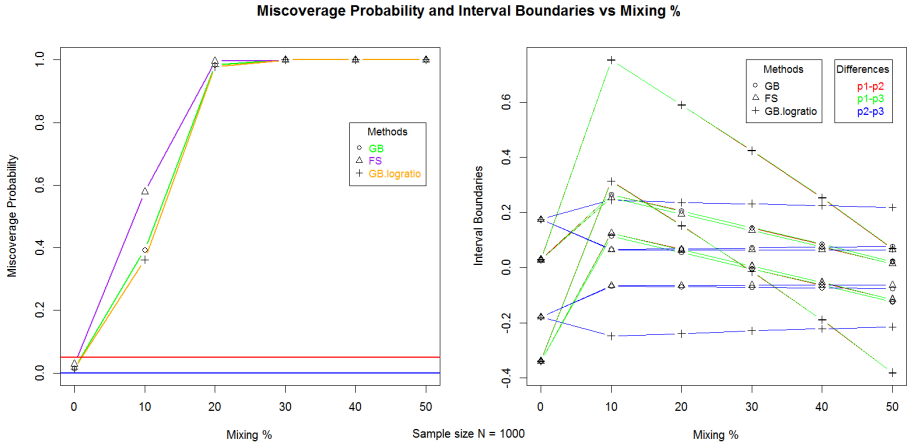

Figure 16: Miscoverage Probability vs Mixing rate with improvement for N=1000

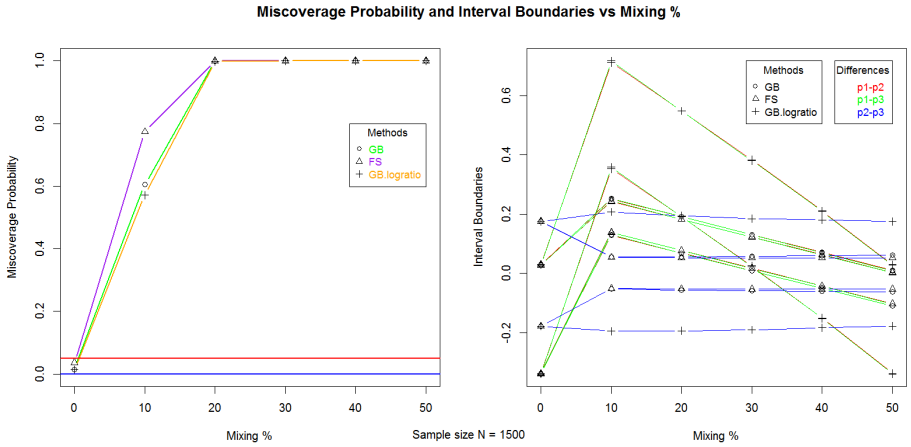

Figure 17: Miscoverage Probability vs Mixing rate with improvement for N=1500

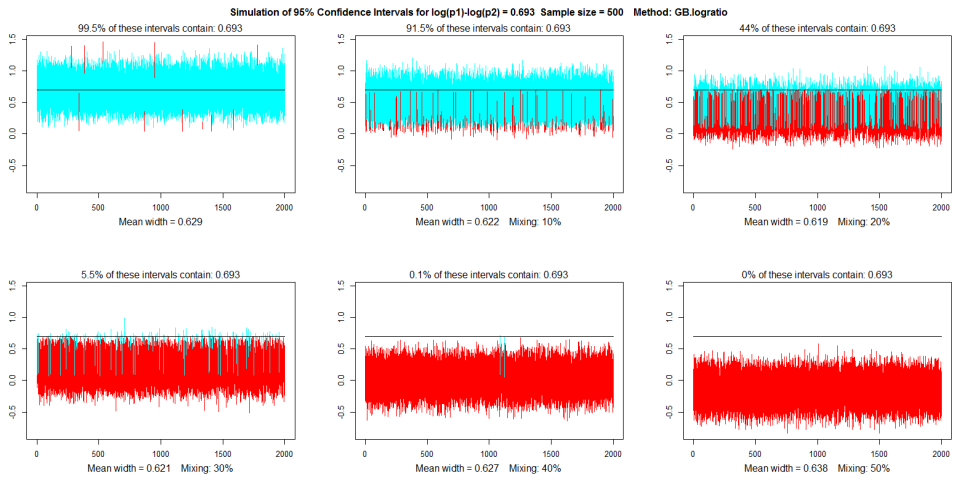

Figure 18: GB Logratio Confidence interval for p_1-p_2 for N=500(PART 2)

8 Appendix: Tables

Table 5: Dispersion of width of CIs related to Part 1 from Method GB

Difference	Sample size	mean_CI	sd_CI	IQR_CI	MAD_CI
p1-p2_part-a	500	0.2227	0.0033	0.0049	0.0036
p1-p3_part-a	500	0.1996	0.0042	0.0061	0.0045
p2-p3_part-a	500	0.1786	0.0044	0.0058	0.0043
p1-p2_part-a	1000	0.1577	0.0016	0.0022	0.0016
p1-p3_part-a	1000	0.1412	0.002	0.0027	0.002
p2-p3_part-a	1000	0.1266	0.0021	0.0028	0.0021
p1-p2_part-a	1500	0.1288	0.0011	0.0015	0.0011
p1-p3_part-a	1500	0.1154	0.0014	0.0019	0.0014
p2-p3_part-a	1500	0.1033	0.0013	0.0018	0.0013
p1-p2_part-b	500	0.253	9e-04	0.0011	8e-04
p1-p3_part-b	500	0.1377	0.0028	0.0038	0.0028
p2-p3_part-b	500	0.1373	0.0027	0.0038	0.0028
p1-p2_partb	1000	0.179	4e-04	6e-04	4e-04
p1-p3_part-b	1000	0.0973	0.0014	0.002	0.0015
p2-p3_part-b	1000	0.0971	0.0014	0.0019	0.0014
p1-p2_part-b	1500	0.1462	3e-04	4e-04	3e-04
p1-p3_part-b	1500	0.0795	0.001	0.0014	0.001
p2-p3_part-b	1500	0.0793	0.001	0.0013	0.001

Table 6: Dispersion of width of CIs related to Part 1 from Method FS

Difference	Sample sizes	mean_CI	sd_CI	IQR_CI	MAD_CI
p1-p2_part-a	500	0.1821	0	0	0
p1-p3_part-a	500	0.1821	0	0	0
p2-p3_part-a	500	0.1821	0	0	0
p1-p2_part-a	1000	0.1288	0	0	0
p1-p3_part-a	1000	0.1288	0	0	0
p2-p3_part-a	1000	0.1288	0	0	0
p1-p2_part-a	1500	0.1051	0	0	0
p1-p3_part-a	1500	0.1051	0	0	0
p2-p3_part-a	1500	0.1051	0	0	0
p1-p2_part-b	500	0.1821	0	0	0
p1-p3_part-b	500	0.1821	0	0	0
p2-p3_part-b	500	0.1821	0	0	0
p1-p2_part-b	1000	0.1288	0	0	0
p1-p3_part-b	1000	0.1288	0	0	0
p2-p3_part-b	1000	0.1288	0	0	0
p1-p2_part-b	1500	0.1051	0	0	0
p1-p3_part-b	1500	0.1051	0	0	0
p2-p3_part-b	1500	0.1051	0	0	0

Table 7: Dispersion of width of CIs related to Part 2 from Method GB

Difference	Sample size	mean_CI	sd_CI	IQR_CI	MAD_CI
p1-p2	500	0.2131	0.0036	0.005	0.0038
p1-p3	500	0.2128	0.0036	0.0053	0.004
p2-p3	500	0.1876	0.0039	0.0052	0.0038
p1-p2	1000	0.1507	0.0018	0.0023	0.0017
p1-p3	1000	0.1506	0.0018	0.0023	0.0017
p2-p3	1000	0.1329	0.002	0.0025	0.0018
p1-p2	1500	0.1231	0.0012	0.0016	0.0012
p1-p3	1500	0.123	0.0012	0.0016	0.0012
p2-p3	1500	0.1085	0.0012	0.0016	0.0012
p1-p2	500	0.2126	0.0033	0.0043	0.0032
p1-p3	500	0.2131	0.0033	0.0041	0.0031
p2-p3	500	0.1948	0.0037	0.005	0.0037
p1-p2	1000	0.1506	0.0017	0.0023	0.0017
p1-p3	1000	0.1506	0.0017	0.0022	0.0017
p2-p3	1000	0.1377	0.0018	0.0027	0.002
p1-p2	1500	0.123	0.0011	0.0015	0.0011
p1-p3	1500	0.123	0.0011	0.0015	0.0011
p2-p3	1500	0.1125	0.0012	0.0017	0.0013
p1-p2	500	0.2113	0.003	0.0041	0.003
p1-p3	500	0.2119	0.0032	0.0046	0.0033
p2-p3	500	0.2014	0.0035	0.0048	0.0036
p1-p2	1000	0.1497	0.0016	0.0021	0.0016
p1-p3	1000	0.1497	0.0016	0.0022	0.0016
p2-p3	1000	0.1424	0.0018	0.0023	0.0017
p1-p2	1500	0.1222	0.0011	0.0014	0.001
p1-p3	1500	0.1223	0.0011	0.0015	0.0011
p2-p3	1500	0.1163	0.0012	0.0015	0.0012
p1-p2	500	0.2093	0.0034	0.0046	0.0033
p1-p3	500	0.2092	0.0032	0.0041	0.0031
p2-p3	500	0.2077	0.0033	0.0045	0.0033
p1-p2	1000	0.1481	0.0016	0.0022	0.0016
p1-p3	1000	0.1481	0.0016	0.0022	0.0016
p2-p3	1000	0.1468	0.0017	0.0023	0.0018
p1-p2	1500	0.1209	0.0011	0.0015	0.0011
p1-p3	1500	0.1209	0.0011	0.0014	0.0011
p2-p3	1500	0.12	0.0011	0.0014	0.0011
p1-p2	500	0.2057	0.0033	0.0044	0.0033
p1-p3	500	0.2059	0.0035	0.0045	0.0033
p2-p3	500	0.2138	0.0031	0.0042	0.0031
p1-p2	1000	0.1456	0.0017	0.0022	0.0017
p1-p3	1000	0.1455	0.0017	0.0022	0.0016
p2-p3	1000	0.1513	0.0016	0.002	0.0015
p1-p2	1500	0.1188	0.0012	0.0015	0.0011
p1-p3	1500	0.1189	0.0011	0.0015	0.0011
p2-p3	1500	0.1236	0.001	0.0014	0.001

Table 8: Dispersion of width of CIs related to Part 2 from Method FS

Difference	Sample size	mean_CI	sd_CI	IQR_CI	MAD_CI
p1-p2	500	0.1821	0	0	0
p1-p3	500	0.1821	0	0	0
p2-p3	500	0.1821	0	0	0
p1-p2	1000	0.1288	0	0	0
p1-p3	1000	0.1288	0	0	0
p2-p3	1000	0.1288	0	0	0
p1-p2	1500	0.1051	0	0	0
p1-p3	1500	0.1051	0	0	0
p2-p3	1500	0.1051	0	0	0
p1-p2	500	0.1821	0	0	0
p1-p3	500	0.1821	0	0	0
p2-p3	500	0.1821	0	0	0
p1-p2	1000	0.1288	0	0	0
p1-p3	1000	0.1288	0	0	0
p2-p3	1000	0.1288	0	0	0
p1-p2	1500	0.1051	0	0	0
p1-p3	1500	0.1051	0	0	0
p2-p3	1500	0.1051	0	0	0
p1-p2	500	0.1821	0	0	0
p1-p3	500	0.1821	0	0	0
p2-p3	500	0.1821	0	0	0
p1-p2	1000	0.1288	0	0	0
p1-p3	1000	0.1288	0	0	0
p2-p3	1000	0.1288	0	0	0
p1-p2	1500	0.1051	0	0	0
p1-p3	1500	0.1051	0	0	0
p2-p3	1500	0.1051	0	0	0
p1-p2	500	0.1821	0	0	0
p1-p3	500	0.1821	0	0	0
p2-p3	500	0.1821	0	0	0
p1-p2	1000	0.1288	0	0	0
p1-p3	1000	0.1288	0	0	0
p2-p3	1000	0.1288	0	0	0
p1-p2	1500	0.1051	0	0	0
p1-p3	1500	0.1051	0	0	0
p2-p3	1500	0.1051	0	0	0
p1-p2	500	0.1821	0	0	0
p1-p3	500	0.1821	0	0	0
p2-p3	500	0.1821	0	0	0
p1-p2	1000	0.1288	0	0	0
p1-p3	1000	0.1288	0	0	0
p2-p3	1000	0.1288	0	0	0
p1-p2	1500	0.1051	0	0	0
p1-p3	1500	0.1051	0	0	0
p2-p3	1500	0.1051	0	0	0