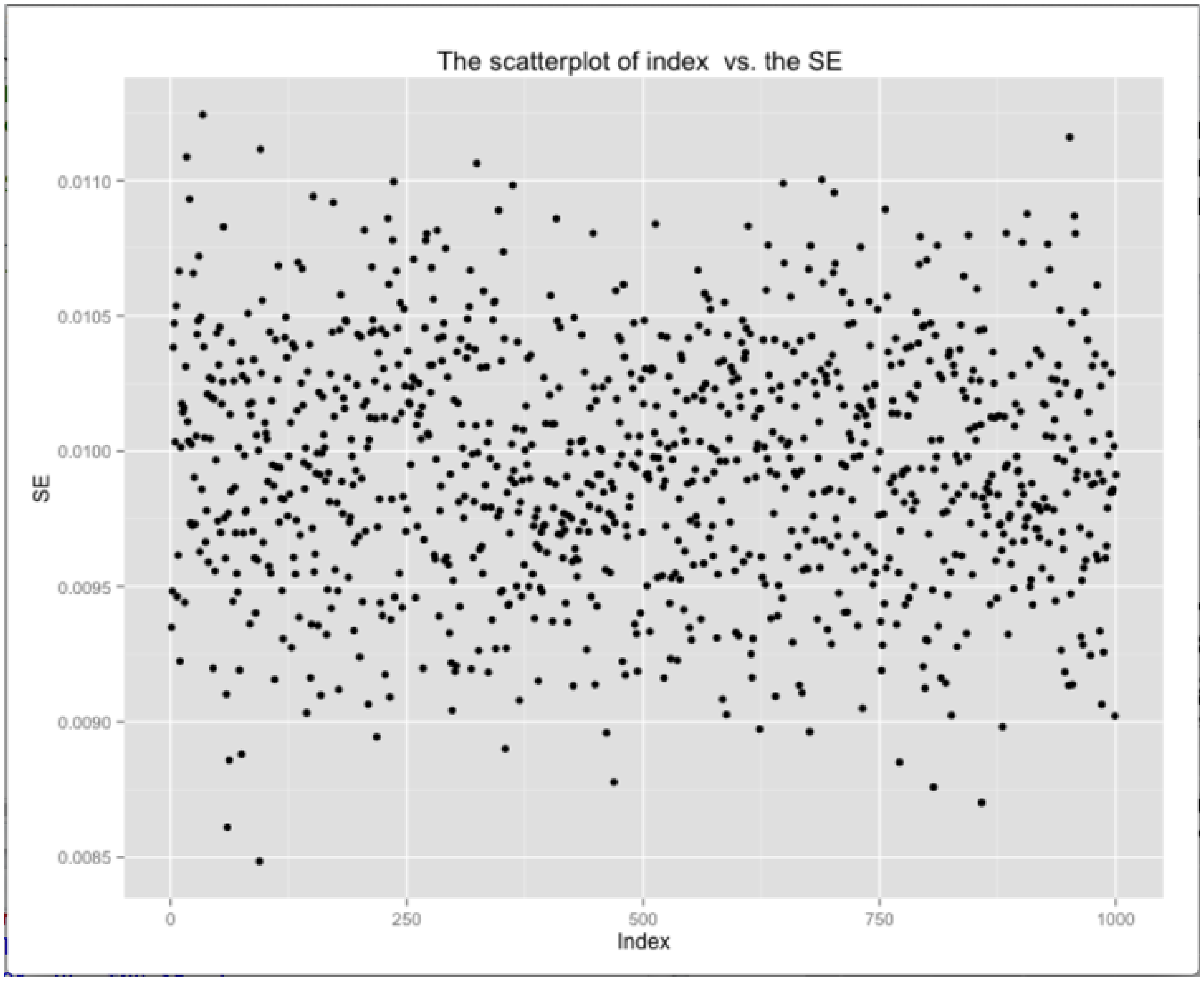
STA 250 Homework2 Qian Li 997838181

Question 1.

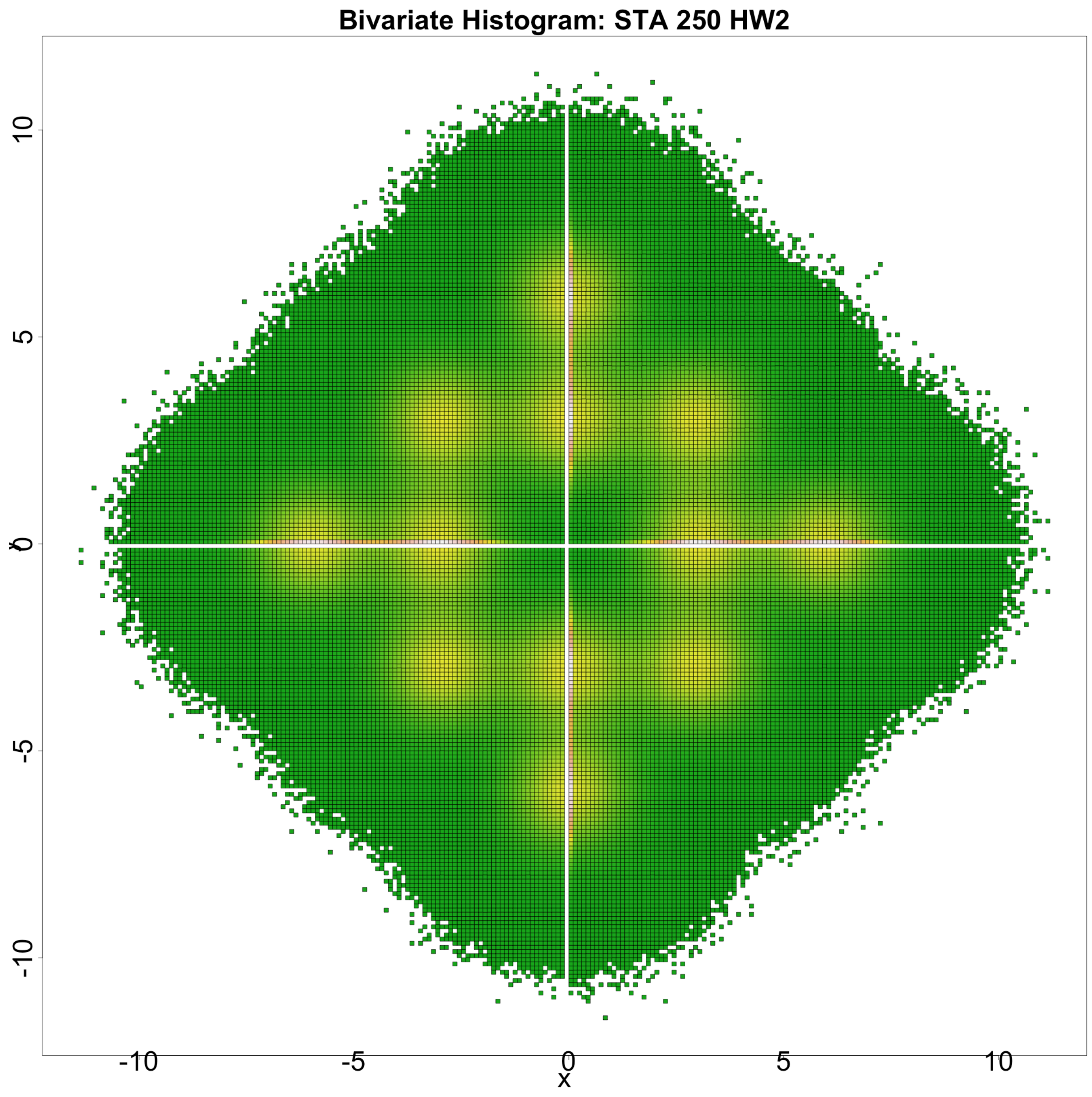


Brief discussion of the algorithm and how well it worked for this example:

1. Randomly sample without replacement 1 subset with b observations from the full data (where b = round(n^gamma); gamma = 0.7 ; n = the number of all observations)
2. From one subset we gain in step 1, we sample with replacement to get one data with n observations. (Using rmultinom)
3. Fit linear model , let the argument weight = the value returned from rmultinom.
4. Repeat step (2, 3) r times, and r = 50.
5. Repeat step (1) s times and s = 5 and step (2, 3) r times, and r = 50.
6. Average over subsamples and Compute BLB SE's and coefficients estimate.

The SE is about 0.01, and I think bootstrap works well in this example.

Question 2.



A description of your MapReduce algorithm. This should address the following questions:

What was your Map function?

1. Transform string to float number
2. Truncate x,y coordinate into bin precision

Keep one decimal for the lower bound of the data, using variable lowerBound
Keep one decimal for the upper bound of the data, using variable upperBound

3. Set x of the bin

For positive number lower bound,
The x_low of the bin = lowerBound, and x_high of the bin = lowerBound+0.1

For negative number lower bound,
The x_low of the bin = lowerBound-0.1, and x_high of the bin = lowerBound

4. Similarly, to set y of the bin

For positive number upper bound,
The y_low of the bin = upperBound, and y_high of the bin = upperBound+0.1

For negative number lower bound,
The y_low of the bin = upperBound-0.1, and y_high of the bin = upperBound

5. Then we change the original data into the form of bins.

What was your Reduce function?

Combine the bins with same x_low, x_high, y_low, y_high together, and count the number of the bins for same coordinate.

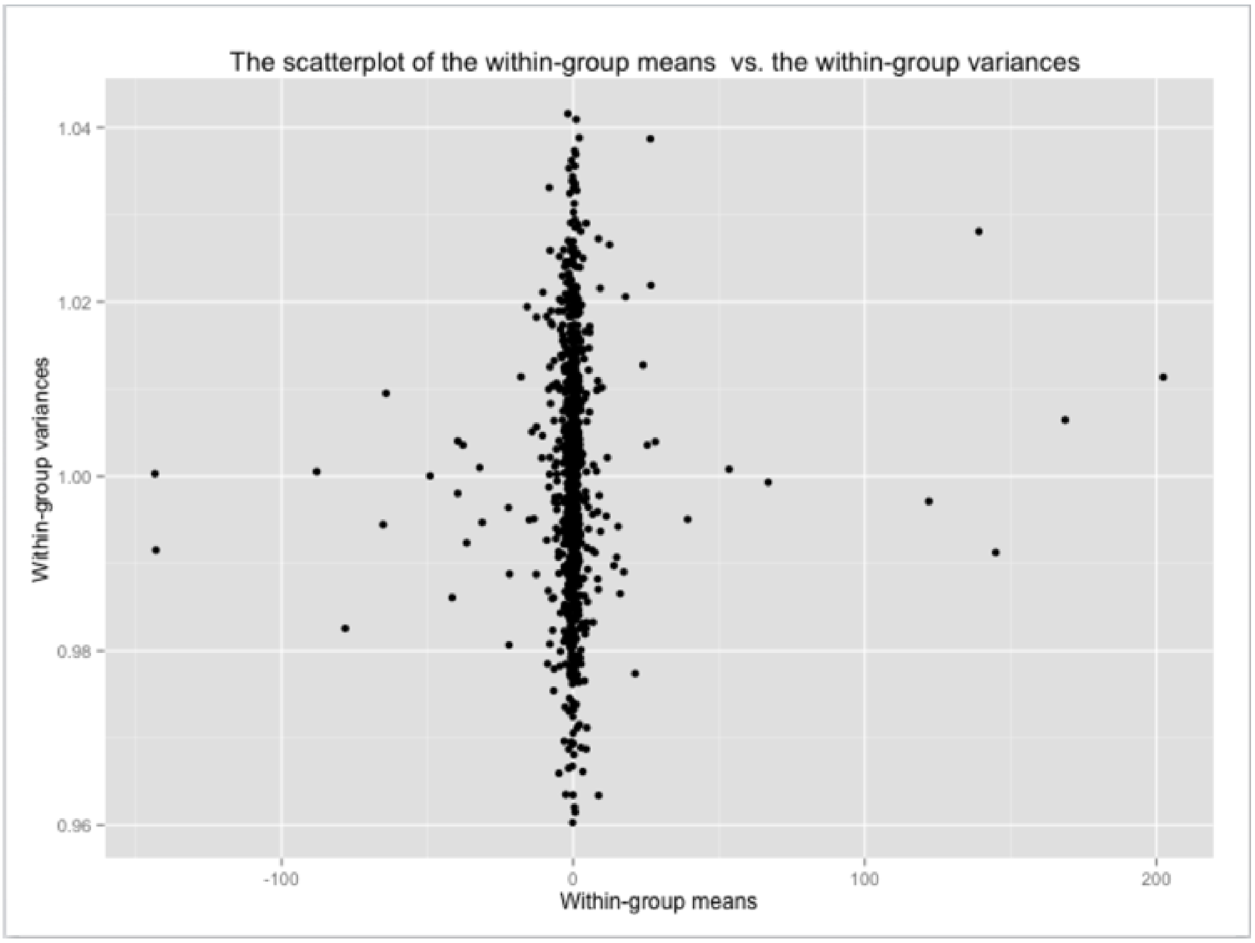
What were the important implementation choices?

How to truncate x,y coordinate into bin precision, and combine same bins rightly.

How well/quickly did your program run?
My code run pretty well and fast, no matter mini data and whole data.

Any comments/improvements you could make?
I think my algorithm for this question is very straightforward and easy.

Question 3.


The scatterplot of the within-group means vs. the within-group variances
Within-group variances
1.04
1.02
1.00
0.98
0.96
-100
0
100
200
Within-group means