STAT8010 Assignment #2

Bernard McNamee

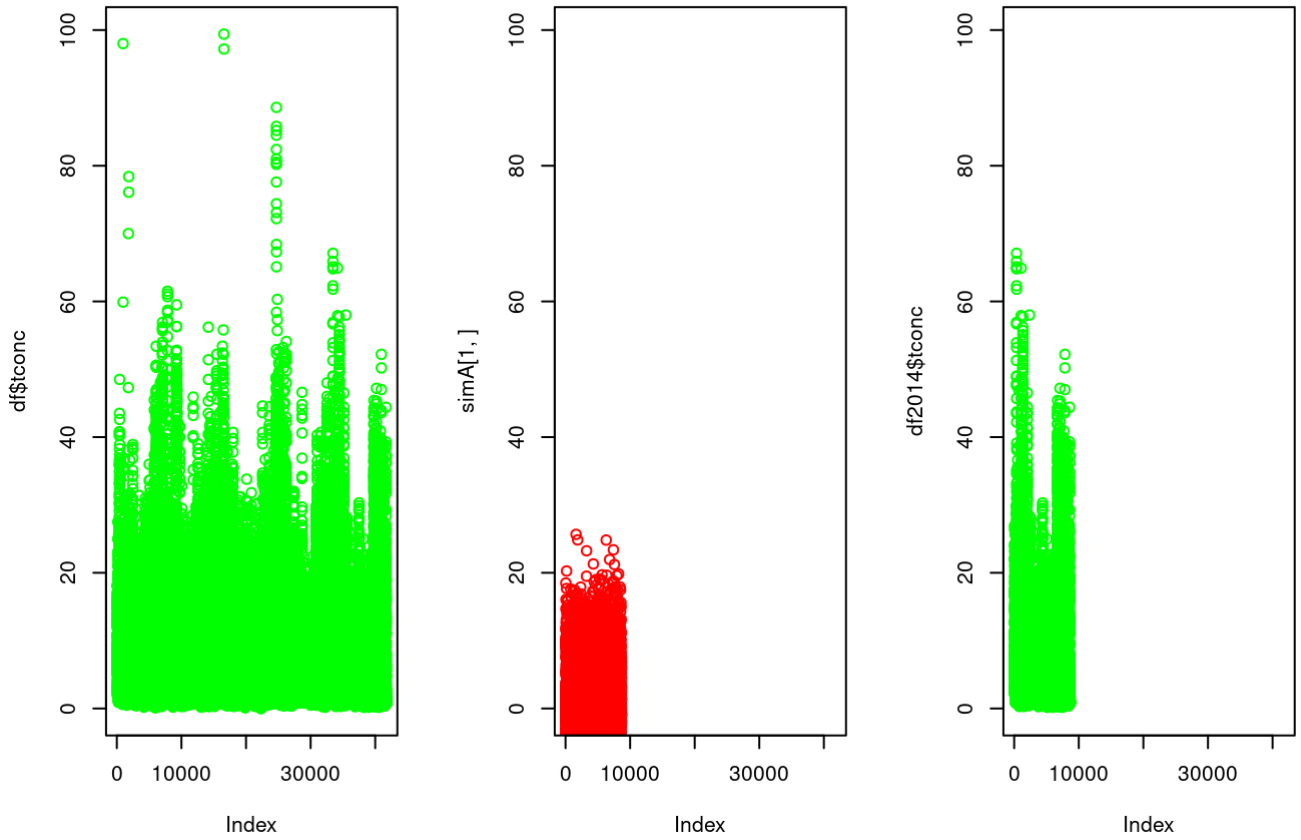
05/01/2021

Answers to questions on various topics - 1 & 2 (Shiny app - see separate R file) and 3, 4 and 5 (Monte Carlo simulations)

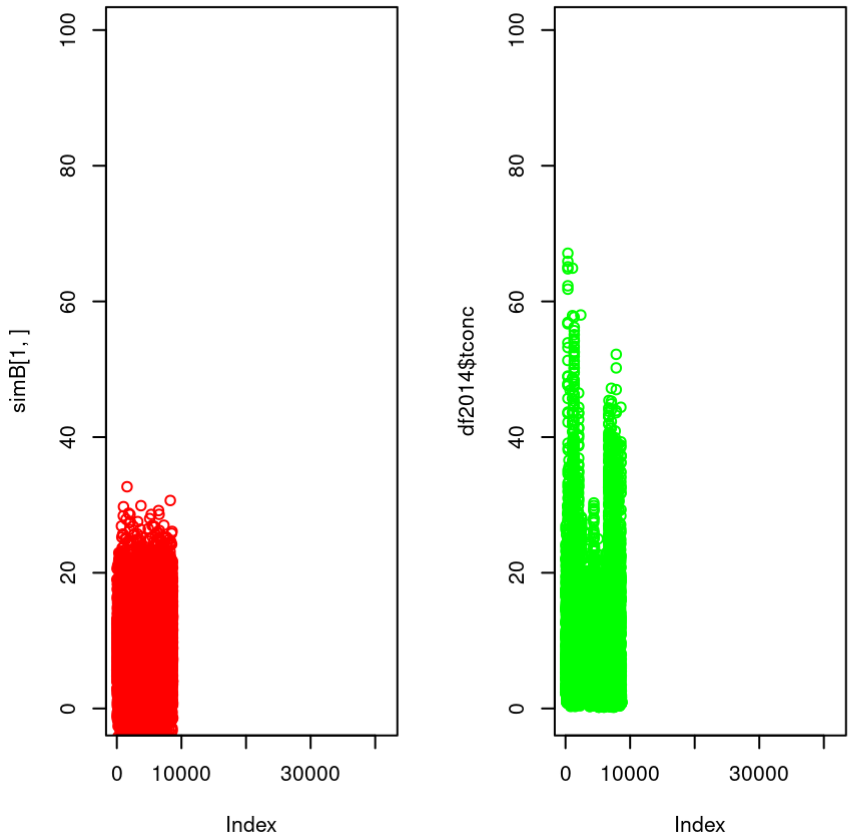
Q3. Using Monte Carlo simulations, you should attempt to predict tconc for the year 2015. This should be done using at least two different models (i.e. different collections of variables or prediction values). You should clearly state which performs best.

The csv data file is read into a R dataframe using base R. The structure is checked and it is noted there is 1. factor variable feed, 2. observation number No, and 3. NA values in tconc. Correlations between variables are checked. A linear regression summary (all variables) shows all p values < 0.05 so all variables are significant, however, the Multiple R squared value is very low indicating a weak relationship between tconc and the variables. A second linear regression summary (excluding time variables bar year) shows again all p values < 0.05, and again the Multiple R squared value is very low (but not much lower without dropped time variables).

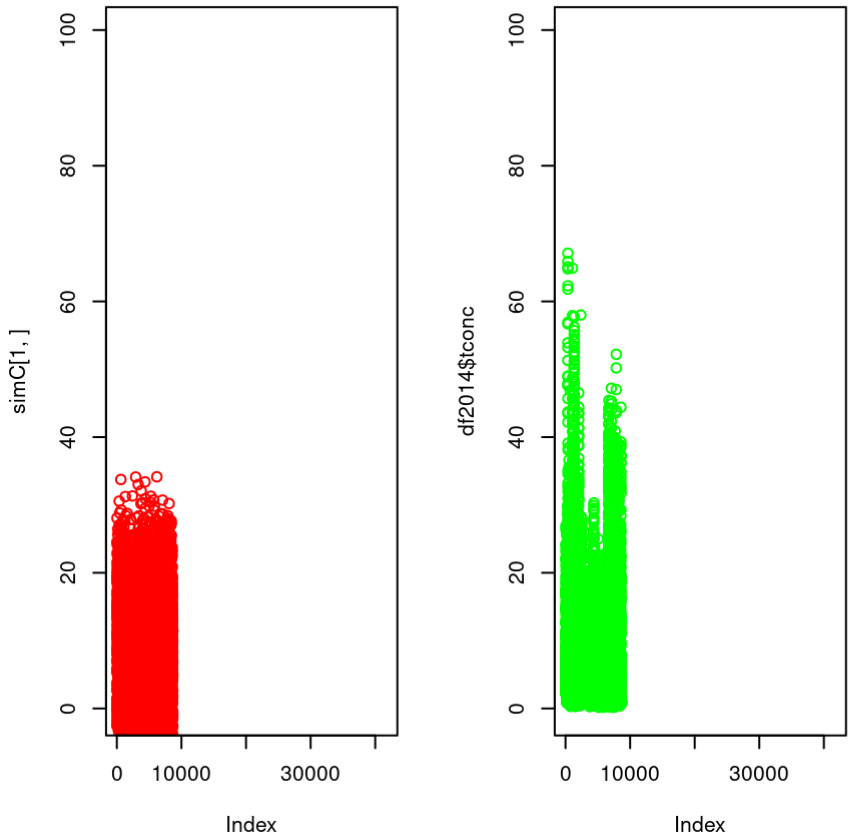
Some simulation setup steps follow; model coefficients (variable coefficients) are calculated and saved; median absolute deviation (instead of sd to avoid outliers) calculated and saved; last row of the main dataframe (simulation starting point) is calculated and saved; number of rows in one year is calculated and saved. Simulation A is run and tconc plotted alongside existing data (for all years and for last year).



A new simulation B is generated using 2014 variable means instead of the last value. Simulation B tconc is plotted alongside existing data (for last year). Unfortunately, the output result appears no better than Simulation A.



A new simulation C is generated using all data from 2014 instead of means / last value. Simulation C tconc is plotted alongside existing data (for last year). Unfortunately, the output result appears no better than Simulation A and B.



Conclusions

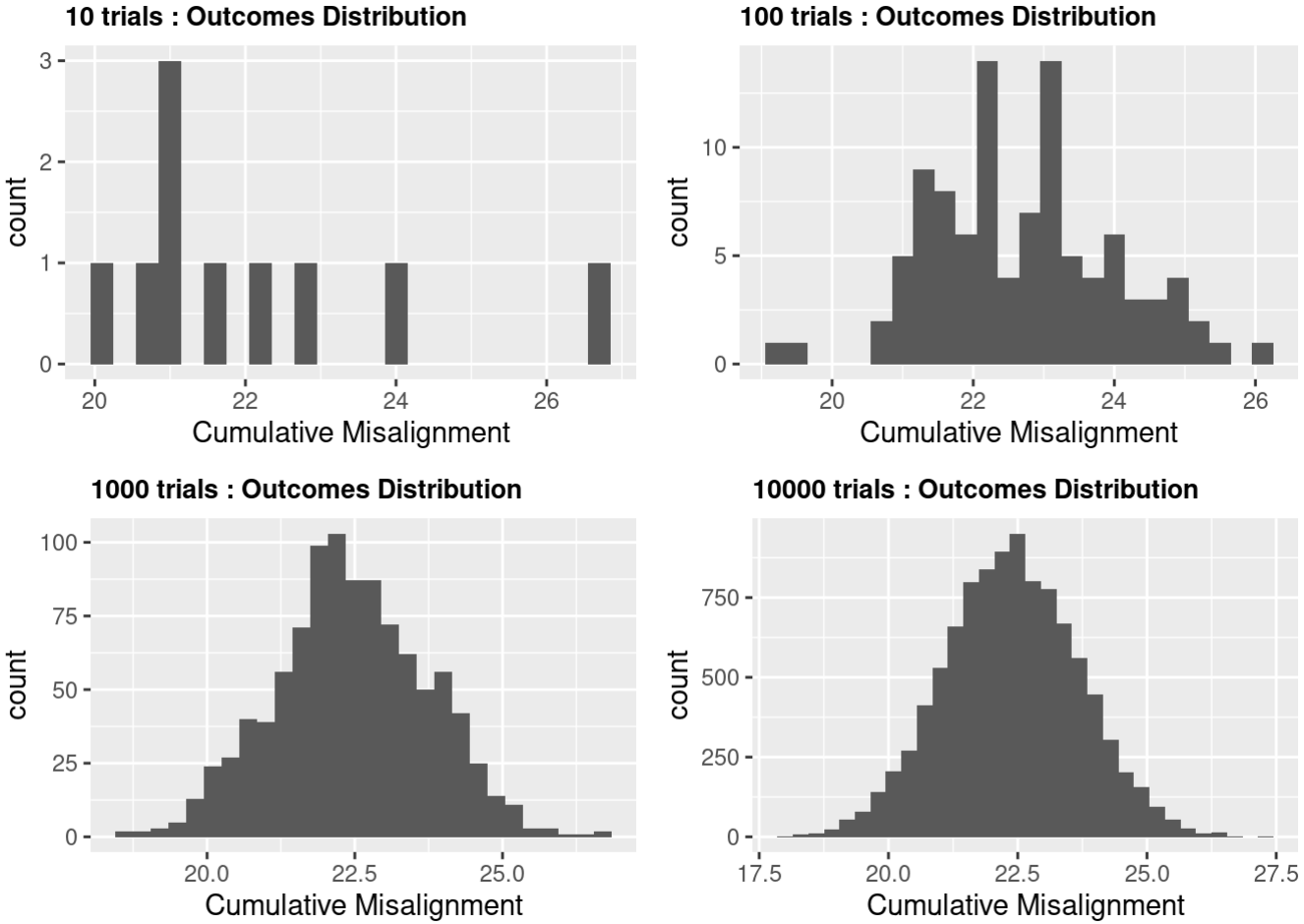
Though weak, there exists a correlation between the simulated data and data from the last year but successive attempts at refining the simulation algorithm did not improve the output result. Even the NA values in the simulated variable were removed and the experiment repeated - no improvement. Although the p values were all much lower than 5%, the Multiple R squared value is very low indicating the relationship between tconc and the dependent variables is weak - not a good starting point. I would have expected Simulation B to give a better result if outliers existed in the starting point values used in Simulation A. And at the very least I would have expected Simulation C to give a better result as it used all values in 2014 for all variables in the simulation algorithm instead of single values (mean/last).

Q4. Consider a machine that inserts a needle into test tubes on a conveyer for sampling in a factory process. This machine may become

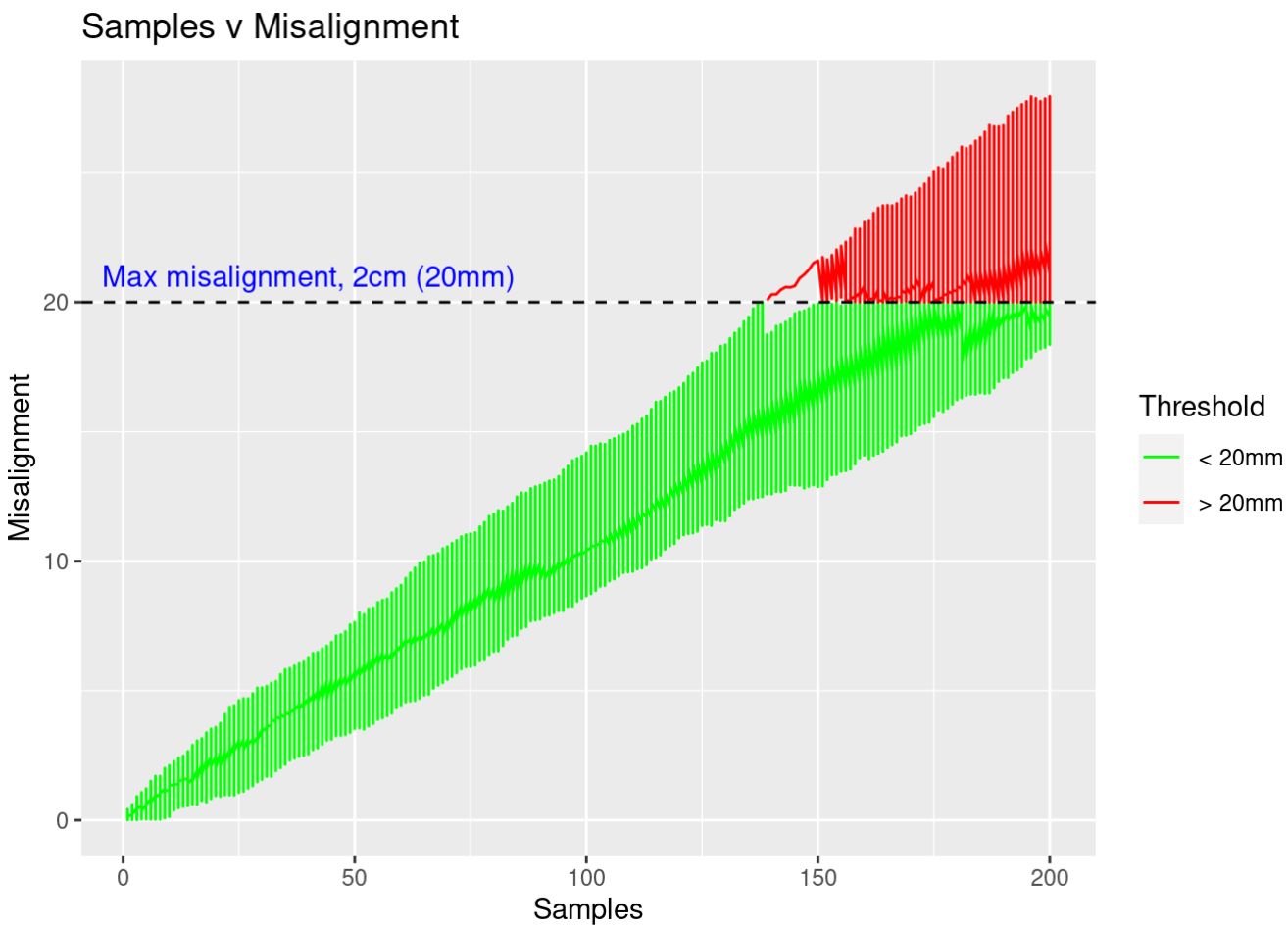
misaligned in the 2 dimensions of the plane of conveyer travel (x and y axes) independently. The machine is realigned to centre at the start of each day and it then samples 200 test tubes throughout the day. The machine fails to sample correctly if it is misaligned in any direction by 2cm or more, as it misses the test tube (possibly colliding with the glass). The x misalignment is 0.1mm on average in the direction of conveyor travel (positive x-direction) for each test, but that this can vary somewhat with a standard deviation of 0.1mm. Similarly, the y misalignment is biased in the negative y-direction, and is much smaller on average; the engineers believe that the average misalignment in the negative y direction is 0.05mm per test, with a standard deviation of 0.05mm.

a. Simulate the distribution of misalignments at the end of the day

As the number of trials increases, the distribution becomes normal as shown below.

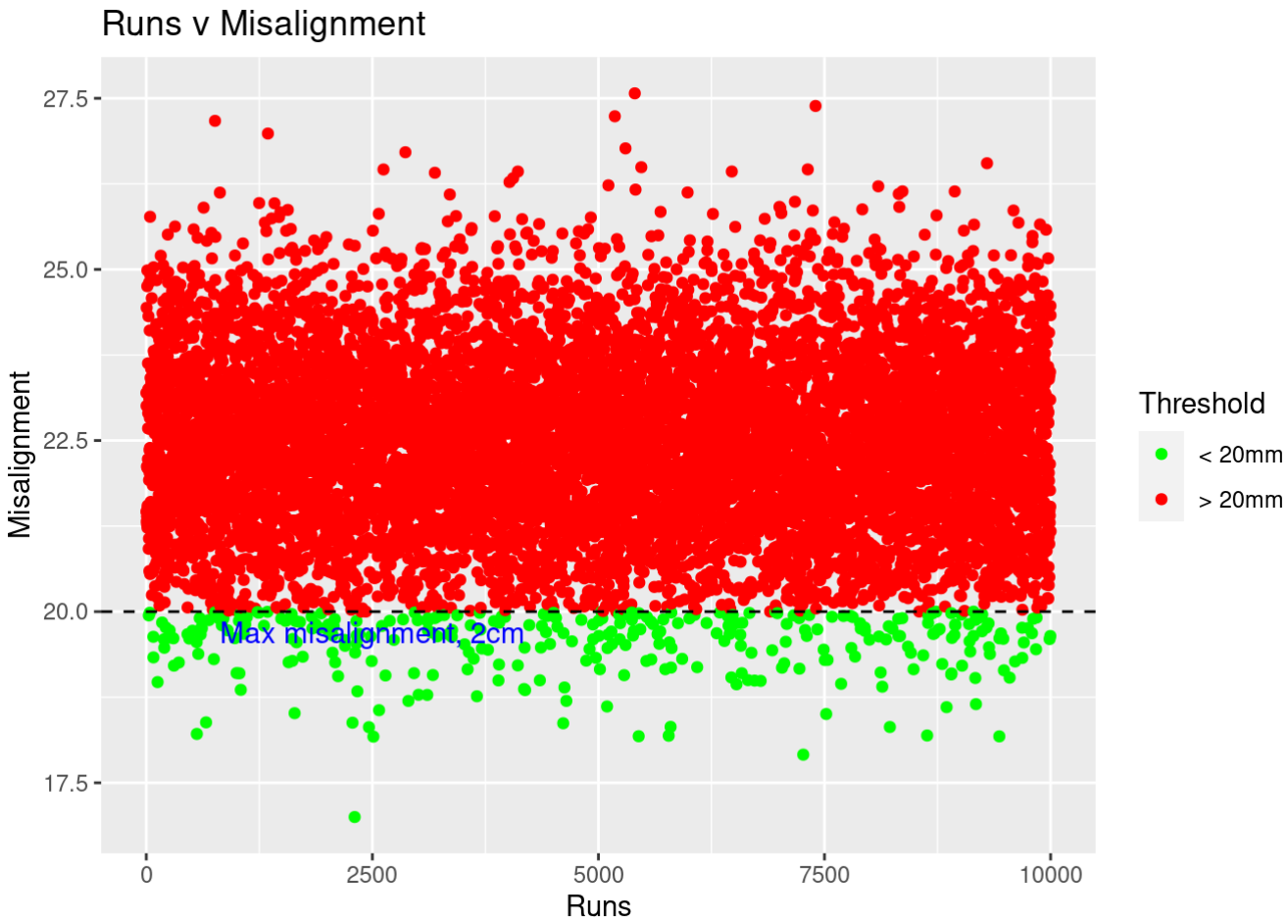


b. Estimate the likelihood of failure throughout the day




```
## [1] "Sum of relative frequencies for large number trial method : Probability of cumulative misalignments > 2cm
(threshold value) at end of a batch of 200 samples is 96.5874 %"
```

c. Visualise the simulated alignments of the machine at the end of the day on a scatterplot, showing the 2cm limit.



Q5. It costs 50,000 when the machine goes offline due to excessive misalignment and no further batches can be tested for the reminder of the day. Each batch passed through the machine results in gross profit of 400. If a batch is ready for testing but the machine is offline, there is a 500 cost for storage and alternate testing of each untested batch under the target number of tests per day. Given these, use Monte Carlo simulations to find the best strategy - i.e. what is the optimal target number of runs per day before realignment should be done.

An out of order sign would be more appropriate - the machine needs to be replaced or repaired. The current batch success rate is so low as to not be worthwhile doing even one batch and risking the offline daily fixed charge for a relatively low return.

Instead, it may be more meaningful to look at reducing the number of samples to a level that will guarantee a batch can complete (see 1% v 5% margin analyses below) before realigning the machine between runs.