– Assignment 1 –

**Q1 (BR)**

*- Briefly describe your simulation process, its goals, and what you have learned from the simulation.*

*- Add at least a plot showcasing the results of the simulation. Make a special note on sample size considerations: how much data do you think you will need?*

*- What else could you do to increase the precision of your estimates?*

Our simulations were designed in order to provide an informed illustration of how our future models should approximately perform. The practical purpose for future modeling would be to predict the number of participants necessary to generate and prove which factors are significant in predicting the linguistic development of autistic children; and further, identify which children would need additional assistance in the classroom.

Chart, line chart

Description automatically generated~~Our simulations ran 10 thousand iterations~~

We ran a simulation of 100 participants with six independent visits each, informed by the prior data of the literature described in the assignment. Considering that our cleaned experiment data is sourced from the performance of 29 ASD and 32 TD participants, the intention of raising the simulated population sample to 100 was to test the diagnostic precision of the prior data for our estimations and modeling strategy going forward.

~~This simulated increase in the population sample also allows tested the precision of our estimations and modeling strategy going forward~~

**Q2 (SS)**

*- Briefly describe the empirical data and how they compare to what you learned from the simulation (what can you learn from them?).*

During the data cleaning process, certain data points were removed, therefore not all participants provide an equal number of data points. This is in accordance with the realistic expectation of children not being equally communicative at all times, or having too much noise in the data, etc. (For instance the child with ID 1 only produced one data point that could be included in the final dataset).

In the raw data there were

The data was recorded throughout the span of 6 visits, where parent-child interactions were recorded and later analyzed.

As a result of cleaning, we got an R tibble containing 352 observations of 20 variables. Participants were anonymized via numeric IDs. 29 of the children were diagnosed with Autism Spectrum Disorder (coded from now on as ASD), and 32 are typically developing (TD).

The variables describe the demographics of each participant, their diagnosis, measures of both verbal and non-verbal IQ, their and their respective mothers’ MLU (mean length of utterance, throughout the visits,

*What did we learn from the simulated data?*

In contrast to the empirical data that contained incomplete data sets for some child participants, the simulated data contains the full 6 data points per participant. Values were randomly sampled from a distribution created based on values (mean and sd)

*- Briefly describe your model(s) and model quality. Report the findings: how does development differ between autistic and neurotypical children (N.B. remember to report both population and individual level findings)?*

*- which additional factors should be included in the model? Add at least one plot showcasing your findings.*

**Q3**

*- Given the model(s) from Q2, how well do they predict the data?*

*Discuss both in terms of absolute error in training vs testing; and in terms of characterizing the new kids' language development as typical or in need of support.*