ARI Sample Size and Power R-Shiny User Guide

Before using the R-Shiny application, the user must determine the following metrics:

1. The type of analysis must be known. In this case there are two options that appear in a drop down:
   * 1. Determine Necessary Sample Size: This allows the user to determine the minimum necessary sample size to achieve at least a specified power for a given hypothesis test
        + - In this analysis, the minimum required power must be known. The default is set to 80% power.
          - The starting value of the sample size, *n*, must be set
          - The value to increase *n* by until the desired power is achieved
     2. Determine Power: This allows the user to determine the power of a specified hypothesis test for a given sample size
        + - In this analysis, the sample size must be known. The default is set to *n=50*.
2. The number of labels for both cluster 1 and cluster 2

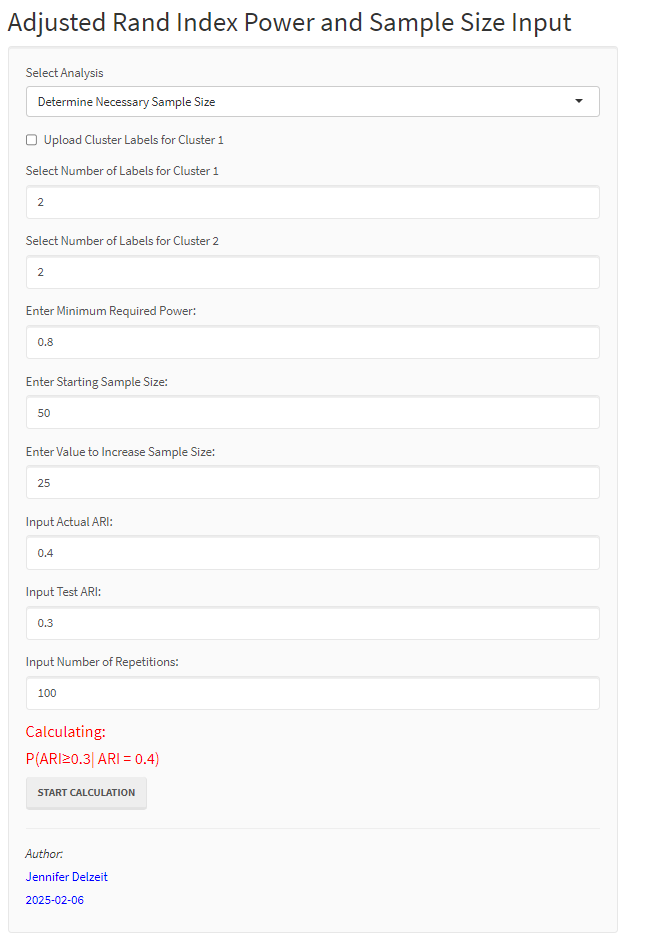
* If the user wishes to input clustering labels for cluster 1, like in the case where a golden standard exists or wanting to compare clustering labels to achieve a specific target, those may be uploaded in a .csv format and the column indicating the labels may be selected after checking the box for “Upload Cluster Labels for Cluster 1”.

\*\*\* It is important to note that uploading clustering labels allows for the user to control the proportions of the item labels in cluster 1. For example, if you upload a dataset that contains 10 rows where your label column has 5 A’s, 3 B’s, and 2 C’s, but input a different sample size other than 10, the algorithm will be forced to have roughly 50% for item label 1, 30% for item label 2, and 20% for item label 3 for clustering 1.

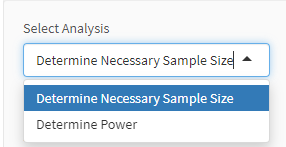
1. The true value of the Adjusted Rand Index (ARI)
2. The value of the Adjusted Rand Index under the alternative hypothesis. In other words, the value of the ARI to be tested against. This value must be at least 0.05 under the true value of the ARI.
3. The number of repetitions the algorithm should run.

**Using the R-Shiny Application:**

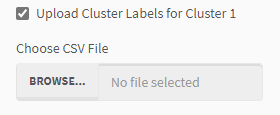
The application appears as follows:



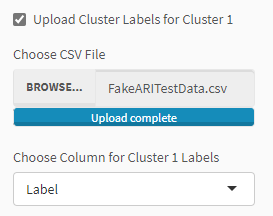
Step One is to set the analysis type from the dropdown:



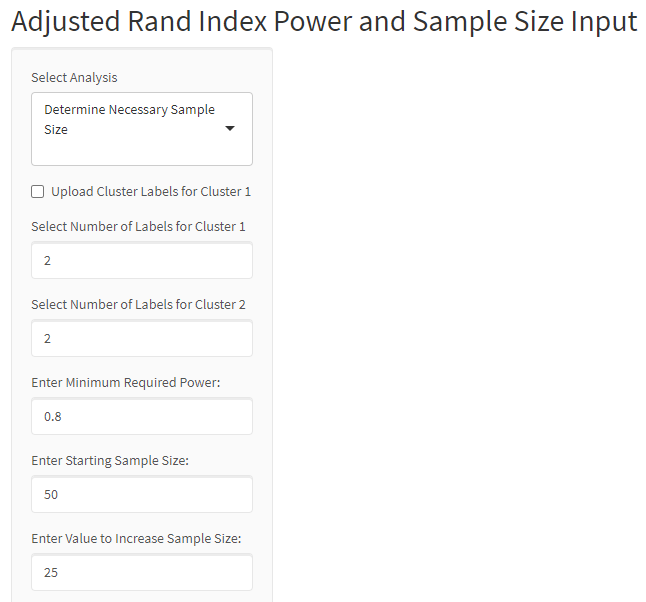
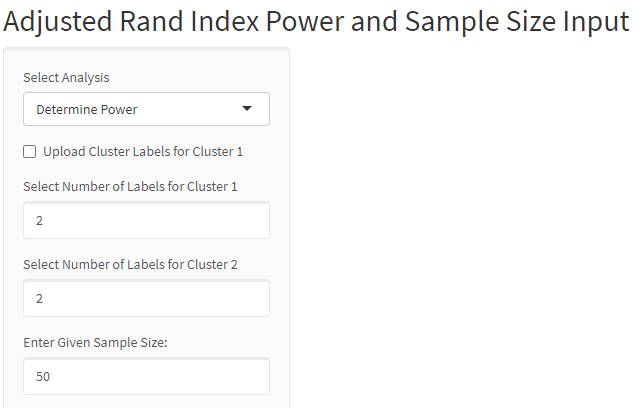
Step Two is to input the desired number of labels for both cluster one and two. If the user wishes to upload a csv file where a column represents the labels for cluster one, the check box next to Upload “Cluster Labels for Cluster 1” may be checked. Upon checking that box, a prompt will appear below to choose the CSV File:



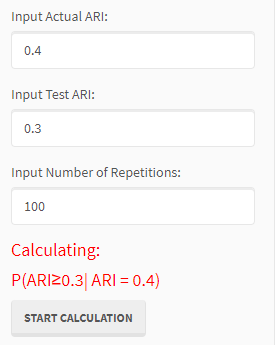
Once the CSV file is chosen and the upload is complete, another drop down will appear to select the column that represents the cluster 1 labels:



After the number of labels for both clusters have been input, the next step is to either input the minimum required power as is the case when the analysis selected is “Determine Necessary Sample Size” or the desired sample size if the analysis selected was “Determine Power”

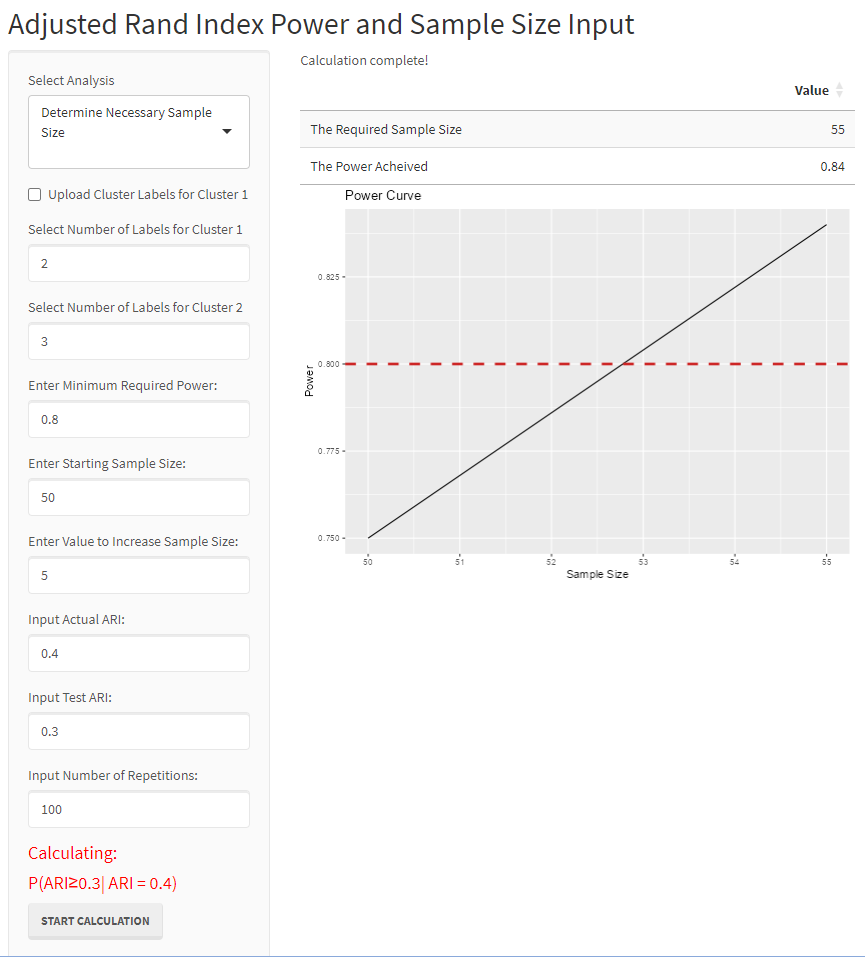


Finally, the last three options are to input the true value of the ARI, the test value of the ARI, and the number of repetitions the simulation should run:



When all inputs have been selected, clicking the “Start Calculation” button will trigger the application to begin running. Once the calculation is complete, a table will appear on the right-hand side. Additionally, the “Determine Necessary Sample Size” option will produce the resulting power curve for the sample sizes that were tested until the required power was achieved.

An example of the table output when the analysis is “Determine Necessary Sample Size” is below. From here, it was determined that the required sample size to achieve at least 80% power was *n=55* where the actual power achieved was 84%.



An example of the table output when the analysis is “Determine Power” is below. From here we glean that the power of this particular test is 88.4%.

