A Muddled Data Generation Process

Hypothesizing and Understanding the Task

Before simulating, I wanted to gain a better understanding of the question and how the distribution would be manipulated by the changes made in steps 1-3. To do so, I sketched what the distribution would look like before and throughout each change. This sketch can be found at Sketches.png.

Simulating the Data

To simulate the data, I first created a normal distribution with 900 observations and a mean and standard deviation of one by using rnorm(). Next, I needed to duplicate the first 100 observations and duplicate them. To do so, I created a variable of the first 100 observations and created a completed sample of the initially simulated and repeated data called full_sample.

Next, I needed to randomly select half of the negative observations and make them positive. I used ChatGPT to find out how to select all the negative values and this resulted in using which(). Once I had all negative observations, I found the number of these observations using length() and divided that by 2. Next, I randomly sampled half of the number of observations from all of the negative observations and created a variable called sample_obs Finally, I selected sample_obs from the full_sample and used abs() to make them positive.

To change the decimal place on the values between 1 and 1.1, I need to remove 1 from all the values. First, I create a variable that encompasses all the numbers between 1 and 1.1. Then I select these numbers in the full_sample and subtract 1.

Finally, to check the mean and whether it is above 0. I created a variable called mean_cleaned_data which is the mean of the full_sample. Then I checked whether that mean_cleaned_data is above 0.