Neuron Showdown: 2-sample hypothesis testing
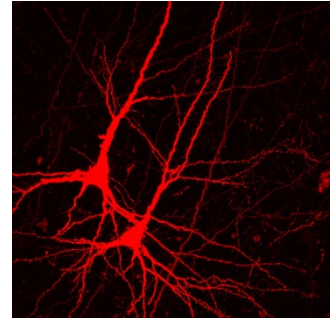
*See MATLAB code in **Perm2Sample.m** for a solution to this problem.*

Learning objectives covered:
- Plot experimental data and outcomes of simulations
- Compute summary statistics from a dataset
- Use computer simulations to build intuition about random variables
- Simulate different kinds of random distributions
- Articulate and code the Null Hypothesis
- Generate a simulated distribution under the Null Hypothesis
- Compare simulated results to data and obtain a p-value
- Compare simulation-based approaches to standard hypothesis tests

We are interested in morphological differences between two different types of pyramidal neurons in the cerebral cortex. One type is called "feedback" (FB) and the other is called "feedforward" (FF). We have used a genetically modified rabies virus to label the two populations, filling each labeled neuron completely. This allows us to reconstruct each neuron and make a number of measurements. For this exercise, will compare the area of the cell body from 50 neurons of each cell type.

Load in the file called 'cellBodyArea.mat'. This file contains two variables, FB and FF, each of which contains fifty values consisting of the measurements of the cell body area for feedback (FB) and feedforward (FF) neurons.

Your goal is to test the hypothesis that the cell bodies of FB neurons are *larger* in area than those of the FF neurons. But instead of using a "classical" statistical test, such as the 2-sample t-test or the Wilcoxon Rank Sum Test, we are going to create our own test and then compare it with the results from classical tests.

Plot a histogram. What is the actual difference in mean values of the two populations? What is the difference in the medians? Remember that our alternative hypothesis is that FB > FF, so we want to keep track of the sign of the difference.

Now we want to ask how likely is it that we got a difference *this big or one that is bigger* under H0? What is the null hypothesis (H0)? There are actually a number of different ways to answer these two questions, but one general approach, known as "the bootstrap," is to simulate running the exact same experiment many times (> 1000) by re-sampling from the original data under the assumption of H0. Imagine this: even if we were sampling from the *exact same distribution*, if we took 50 values and arbitrarily labeled them "FB" and another 50 and arbitrarily label them "FF," we would rarely expect their means to be exactly the same even though on average we might expect them to be relatively close—this variability due purely to our sampling procedure is referred to as "sampling error." But how often would we expect to get a signed difference (FB – FF) as large or larger than the one we got? This is the question you will answer with a bootstrap test.

When you've got a test with which you're happy, use it on your Yeast data from Day 1 (cell sorting, comparing the YFP levels in red cells vs. blue cells). Are the two groups of cells statistically different?

Other things to try:
1. Compare the output of your test to a 2-sample t-test.
2. Convert your script to a function. What parameters do you need to pass? to return?
3. Allow the user to specify a one- vs. two-tailed test.
4. Allow the user to specify different test statistics for comparison (e.g. mean, median, kurtosis)

Revisions:

RTB wrote it, 24 April 2013 as 'CellBodyAreaAnalysis.m'

MIS added Learning Objectives, question about Yeast data, 31 July 2014

RTB changed it to do resampling with replacement (i.e. bootstrap) 13 July 2016 (solution code more general, now called 'Resample2sample.m')

RTB converted it back to a permutation test. Efron: "When there is something to permute, it is a good idea to do so" (solution code 'Perm2Sample.m')

RTB made up new data with a larger sample size (n=50 for each group), 31 July 2018

RTB re-wrote 'Perm2Sample.m' for more flexibility, 31 July 2018

RTB revised to more generally introduce the concept of the bootstrap, 15 August 2020