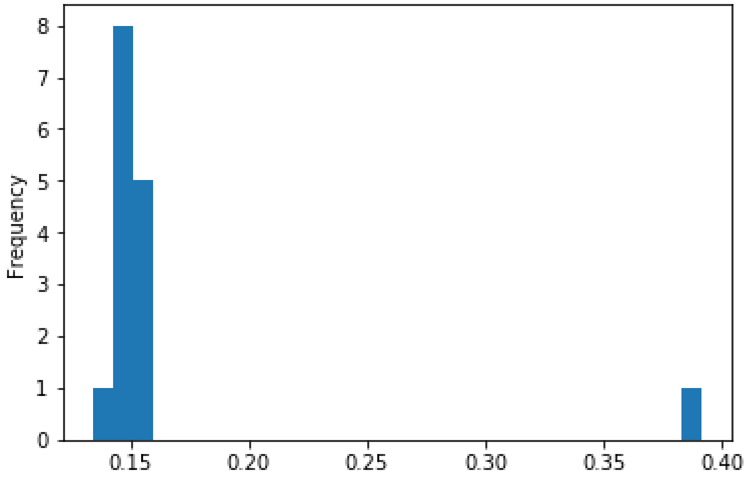
This dataset (https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression) consists of 8 groups (a 2 x 2 x 2 design) with 7-10 biological replicates (i.e. mice) per group and 15 technical replicates (measurements) per mouse. The measurements are of the levels of 77 proteins.

Thus missing data or outliers can be missing or outlying technical replicates or entire missing/outlying biological replicates.

Given the relatively small sample sizes, I chose not to attempt any removal/interpolation of outliers/missing data among biological replicates. However, I felt that extreme outliers among technical replicates was likely to reflect technical error.

I defined an outlier as 3 standard deviations away from the mean within that mouse x protein. Only 70 data points met this criterion.

Example of the data for a single mouse and single protein, showing one extreme outlier:

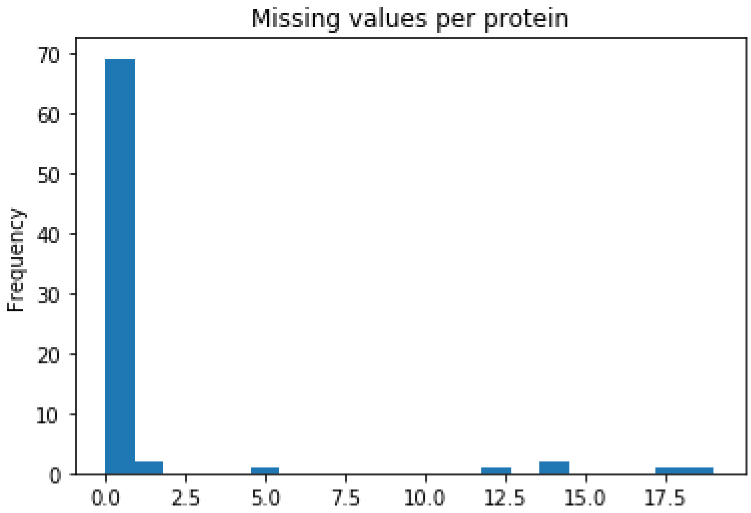


I replaced each of these outliers with null values.

I also found that two of the proteins (‘pS6_N’ and ‘ARC_N’) have duplicate values, so I merged these columns and relabeled it with both names.

I then collapsed the data set by mice by taking the means (and also standard deviations) of each set of technical replicates.

This dataset contains a total of 84 missing values (out of 76 x 72 = 5472 expected values). These are not evenly distributed; for 68/77 proteins, no data is missing.



No group is entirely missing data for a protein; the minimum sample size per group is 5.