Description of the hierarchical bootstrap for the Kaeser lab data.

When the Kaeser lab knocks out two genes involved in synaptic release (RIMs & ELKS), synaptic transmission is reduced by about 80%. Now they want to ask whether KO'ing a 3rd gene, Munc13, further reduces synaptic transmission. However, since the double KO (RIMs/ELKS) is in a different mouse *strain* from the triple KO (RIMs/ELKS/Munc13), the synaptic transmission in each strain's KO must be compared to its own control.

S\_1 = Strain 1: RIMs/ELKS-Cre

S\_2 = Strain 2: RIMs/ELKS/Munc13-Cre

To perform an experiment in a given *strain*, several mice are killed, the hippocampi (HC) are dissected, cells dissociated and pooled together in one big primary culture. This culture is then divided into two pools subjected to different *conditions*: one is treated with a lentivirus containing Cre (= KO); the other is treated with a lentivirus containing a dead Cre, 'delta-Cre' (= Control). From each culture, multiple *cells* are tested. Each cell is patch clamped and synaptic transmission is tested by measuring the size of the post-synaptic current (EPSC or IPSC) evoked by an action potential. This measurement (= *sweep*) is generally repeated 5 times for each cell.

C\_1 = Condition 1: KO

C\_2 = Condition 2: Control

Thus each measurement is uniquely identified by FIVE numbers, which will be variables (columns) in the Excel file. The Excel file will be read in as a Table variable, 'ds':

ds.Strain [1:2]

ds.Cond [1:2]

ds.Batch [1:nBatches], always 3 batches per experiment

ds.Cell [1:nCells], may vary for different batches, varied from 6 to 12 cells per batch

ds.Sweep [1:nSweeps], usually 5 or 6 per cell

For ease of discussion, we'll define four experimental *groups*:

**Group A**: S\_1, C\_1 (RIMS/ELKS KO)

**Group B**: S\_1, C\_2 (RIMS/ELKS Control)

**Group C**: S\_2, C\_1 (RIMS/ELKS/Munc13 KO)

**Group D**: S\_2, C\_2 (RIMS/ELKS/Munc13 Control)

The scientific question is whether knocking out Munc13 (triple mutant, S2) causes a greater relative decrease (vs. control) in synaptic transmission as compared to the double mutant (S1). So we will define our test statistic, *T*, as:

*T* = [mean(Group\_A) / mean(Group\_B)] / [mean(Group\_C) / mean(Group\_D)]

The null value for this statistic is 1, and our alternate hypothesis is:

*T* > 1

In the hierarchical bootstrap (Saravanan et al. 2020), we estimate the sampling distribution of our test statistic, *T*, by resampling *with replacement* from the raw data, while preserving the hierarchical relationships created by the design of the experiment (**figure 1A**). For each bootstrap iteration, *group* identity (combination of *strain* and *condition*) is preserved. We do our resampling at three nested levels: *batch*, *cell* and *sweep* (technical replicates). In the framework of a linear mixed effects model, *group* (= *strain* x *condition*) would be a fixed effect and *batch*, *cell* and *sweep* would be random effects.

For each of the four experimental groups, we perform the following hierarchical resampling. Starting with group A (RIMS/ELKS KO), we first resample *with* replacement from the batches of cells in this experiment. In all cases, there were three batches per experiment, so we would use the MATLAB command 'unidrnd(3,3,1)' to draw 3 random samples from the uniform discrete distribution from 1 to 3. We might, for example, draw B2, B2, B1 for this iteration (as in the example in **figure 1B**). We would then start with batch 2 and determine the number of cells sampled in this batch and randomly resample from the cells, replicating the same number of cells but containing a different combination from that of the actual experiment. If there were 7 cells tested in batch 2, we might draw the following bootstrap sample: C1, C3, C7, C6, C7, C5, C6, in which cells 6 and 7 are included twice and cells 2, 3 and 4 not at all. Then we would proceed through this list of cells, each time randomly resampling (always *with* replacement) from the set of technical replicates for that cell and appending these resampled measurements to the bootstrap sample for this group. After doing this for each of the 7 cells, we would go back and repeat the entire process, first for batch 2 (again), but selecting a different sample of the 7 cells, and then for batch 1. At the end we would have a bootstrap sample for this experimental group that was exactly the same size as our original data set, but it would contain a different subset of measurements and, importantly, it would preserve the nested, hierarchical structure of our experiment. We would repeat the above process for the remaining three groups (B, C and D), then calculate *T\** using the formula for *T* above (The '\*' denotes a bootstrap replicate of our test statistic.) This entire procedure was repeated 100,000 times, producing an estimate of the sampling distribution of *T\**. Based on the distribution of the 100,000 *T\** values, we calculated 95% confidence intervals using the percentile method (**figure 2**). In addition, we directly calculated the probability of H0 given our data as:

p(H0|data) = #{*T\** ≤ 1} / 100,000

Note that this is not the traditional p-value calculated with standard statistical tests, which is the probability of obtaining a result as extreme or more extreme *assuming H0 to be true*. Rather the metric computed above is a more direct measure of a given hypothesis being true given our data (Saravanan et al. 2020). Yet another suggested interpretation of this metric is as the "Type S error," which is the probability of the measured effect being in the wrong direction (Gelman & Carlin 2014). In this case, values of *T* less than 1 correspond to the synaptic transmission in the Munc13 KOs being greater than that in the double KOs. Finally, a more traditional form of frequentist inference can be performed using the bootstrap confidence interval: if the null value (*T* = 1) lies outside of the 95% confidence interval, then H0 can be rejected at a significance level of α = 0.05 (Efron & Tibshirani 1994). Regardless of the exact procedure used, our data are consistent with a highly significant effect of the RIMS/ELKS/Munc13 manipulation in the direction of *decreased* synaptic transmission as compared to that for the RIMS/ELKS KO.

It seems *a priori* plausible that the design of the experiments would impart structure to our data. That is, it is likely that repeated measurements (*sweeps*) from one *cell* would be more similar to each other than to measurements from another *cell*. And we might also expect that the cells from one *batch* would be more similar to each other than to those from other *batches*, despite our best efforts to standardize the treatment of different batches. But is this in fact the case? Might it still be legitimate to treat all of the measurements from a given experimental group as independent?

To explore this question, we performed two additional analyses. First, within each experimental group (i.e. eliminating the fixed effects of *strain* and *condition*), we performed 2-way ANOVA with *batch* and *cell* as factors. For the four experiments in which there were repeated measurements for single cells (i.e. excluding the 2 sucrose experiments), we ran the ANOVA for each of the four groups, yielding 16 tests, and found that *cell* was a significant factor in 14 of them, with a criterion of α < 0.05, 11 of which were significant at α < 0.001. Including all six experiments, we found *batch* to be a significant factor less frequently, but there was still evidence for higher between-*batch* variance than within-*batch* variance in 50% of the 24 cases. This is perhaps not so surprising, because each batch represented the admixture of neurons from several animals and were thus likely to be more homogeneous. Nevertheless, the ANOVA provides strong evidence that we were right to be concerned about structure in our data.

The second analysis was to repeat the bootstrap in a non-hierarchical way and compare the distributions of *T\** and the resulting confidence intervals with those from the hierarchical bootstrap. To do this, we simply pooled the measurements from all of the batches and cells from a given experimental group, and resampled with replacement from this pool. This instantiates the assumption that all measurements were independent. As expected, these distributions were much narrower and more bell-shaped than the corresponding distributions from the hierarchical bootstrap (**figure 2**), with the resulting 95% confidence intervals being, on average, less than half the width of their hierarchical counterparts. Importantly, extreme values of *T\** were much less likely to be represented in the non-hierarchical distribution, and the significance level of the results was overestimated. While this analysis does not directly address the issue of structure in the way that the ANOVA does, it provides a vivid picture of the consequences of *not* analyzing the data in the appropriate way.

References:

Efron B and Tibshirani RJ (1994) An introduction to the bootstrap. CRC press.

Gelman A and Carlin, J (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors." Perspectives on Psychological Science, 9(6), pp.641-651.

Saravanan V, Berman GJ, Sober SJ (2020) "Application of the hierarchical bootstrap to multi-level data in neuroscience." *Neuron Behav Data Anal Theory*. 3(5).

https://nbdt.scholasticahq.com/article/13927-application-of-the-hierarchical-bootstrap-to-multi-level-data-in-neuroscience. Epub 2020 Jul 21. PMID: 33644783; PMCID: PMC7906290.