Soundness of bootstrap resampling procedure

We report here the results of simulations that indicate the soundness of our bootstrapping procedure. These simulations also justify the soundness of the analysis in Lorca-Puls et al (2018, Neuropsychologia). The code of these simulations is available at the online repository for this study.

We assume a "base-sample", which in this paper would be the collection of 313 participants that performed each of the four executive function and implicit learning experiments. For Lorca-Puls et al it would be a large set (360) of patients selected from the Ploras database according to some criteria and for the simulations discussed here a large set (360) of numbers sampled from a Gaussian.

The key property we are seeking to verify is the following:

> *For a given "base-sample", sampling directly with replacement (direct_wi_rep) from the base-sample generates the same uncertainty (i.e. standard deviations (stds)) as sampling indirectly with replacement (indirect_wi_rep) from the base-sample.*

The simulations proceed as follows, using the sample sizes employed in Lorca-Puls et al.

Basic procedure

1) Sample a base sample of size 360 from a gaussian, call this *base-sample*.
2) Sample from *base-sample* to generate subsamples with sizes ranging from 330 to 30 (in steps of 30), we denote these sizes with M. For all of these subsample sizes, N samples are taken. Thus, we are left with N samples all of size 330, N samples all of size 300, N samples all of size 270, and so on down to N samples all of size 30. (Conceptually, one of the N samples of a particular size represents the data that could have been collected in a single experiment with M participants.)
3) These subsamples were taken under four different sampling procedures,
    a. *direct_wi_rep*: this is a bootstrapping procedure (hence, *wi_rep*, i.e. with replacement), and the bootstrap samples are taken directly from the *base-sample* (hence, *direct*).
    b. *direct_w/o_rep*: this again samples directly from the *base-sample*, but it is not bootstrapping, since samples are taken without replacement; as a result, there is no duplication of elements within each sample, although different samples can include the same elements.
    c. *indirect_wi_rep*: this is a bootstrapping procedure, but it involves two levels of sampling; that is, a first sample is taken (without replacement) from the *base-sample*, which we call the *intermediate* sample, and then subsamples proper are bootstrapped from this *intermediate* sample.
    d. *disjoint*: this involves performing disjoint splits of the *base-sample*; accordingly, the smallest samples that can be generated are of size 180 for a *base-sample* of size 360

    We also have a *reference* condition, which is a ground truth and involves sampling directly from a gaussian, rather than directly or indirectly from a *base-sample*.

4) For any subsample size and sampling procedure, we proceed as follows,
    a. We will have N subsamples.
    b. Calculate the Cohen's d (difference from zero) effect size of each of these N subsamples.

c. The resulting (N item) distribution of Cohen's d's is playing the role of one of the distributions of effect sizes in the main body of this paper.
d. We then calculate the standard deviation of this distribution of Cohen's d's, which plays the role of the dispersion of one of these distributions of effect sizes. We call this the *basic procedure standard deviation*.

We now iterate over this basic procedure in order to determine the statistics (central tendency and dispersion) of the *basic procedure standard deviations*. Thus, we run the basic procedure many times and calculate a mean estimate of the standard deviation of basic procedure distributions. We also calculate the standard deviation across these *basic procedure standard deviations*, in order to understand the variability in a single running of the basic procedure.

Results

The results of our simulations are shown in figures 1 and 2.

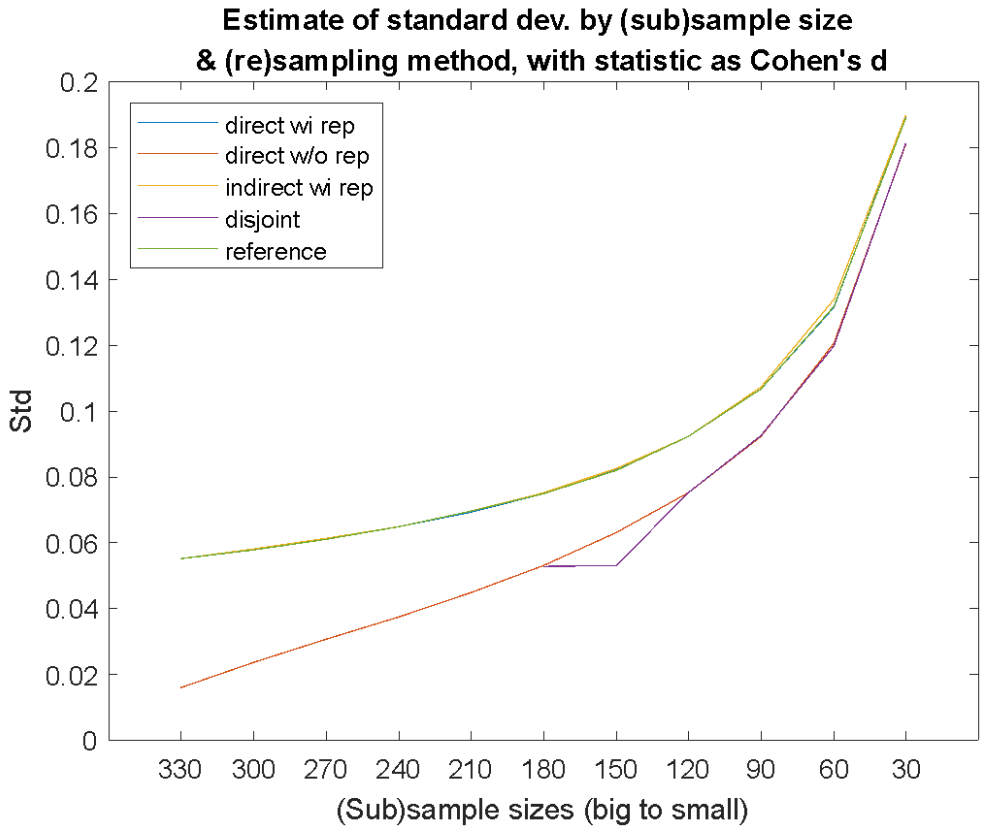

Figure 1: the mean estimate of basic procedure standard deviations, when the "main statistic" is the effect size. This corresponds to a central tendency estimate of the standard deviations of basic procedure distributions. This is given for a range of sub-sample sizes, and plotted for the four different resampling procedures, with the nature of the disjoint split procedure meaning that it cannot have subsample sizes greater than 180. The "ground truth" *reference* procedure is also shown. The lines for *direct_wi_rep*, *indirect_wi_rep* and *reference*, are sat on top of each other, while *direct_w/o_rep* and *disjoint*, are sat on top of each other. Importantly, *direct_wi_rep* and *indirect_wi_rep* show the same pattern, which is our finding.

Figure 1 is the main finding. Central tendencies are being estimated of standard deviations of sets of parameter estimates (here, effect sizes), each of which was generated from a resampling. So, this mirrors what happens in the main body of this paper, where bootstrap samples are generated, effect sizes of each are calculated and the results are put into a distribution.

Critically, in figure 1, the *direct_wi_rep* and *indirect_wi_rep* lines track each other. *direct_wi_rep* is the procedure employed in the main-body of this paper and Lorca-Puls et al, i.e. bootstrap smaller and smaller samples from single (eventually much bigger) root data set (our *base-sample*). Consequently, the smaller samples will have less items in common between samples, than the larger samples, and one might think this is what causes the increased variability we see as samples get smaller, i.e. the increase in curves as one moves from left (large samples) to right (small samples) in figure 1.

The *indirect_wi_rep* line counters this, since it first samples an "intermediate" sample from the large, base, sample, which is the size of the bootstrap samples then generated from it, e.g. the *indirect_wi_rep* 60 data point is generated from bootstrap samples of size 60 from a sample that was of size 60. Thus, at each sample size, bootstrap samples were generated from samples of the same size as the bootstrap samples.

Importantly, this does not change the pattern we observe of mean estimates of standard deviations. An explanation of why the *indirect_wi_rep* curve is not different from the *direct_wi_rep* one is the following. Even if, on the *direct_wi_rep* curve, if you mapped subsamples to underlying sets (mathematically, sets do not reflect repetition, just containing each item that arises at least once) and then took the intersection, you would find less overlap, as subsamples get smaller, that does not automatically mean the procedure is invalid. That is, the variability in bootstrapping is generated from the varying number of times (including zero times) that the same items appear in a set. The combinatorics of the variability generated in this way is so great that it swamps any decreased overlap that would be apparent when going to underlying sets (i.e. removing repetitions). Indeed, this reflects what is the essential property that makes bootstrapping work. For instance, when bootstrapping samples of size R from a set of size R one will often generate bootstrap samples where the underlying set of items in the samples is the same, but the number of times each item appears in the bootstrap samples varies; it is this variability in number of repetitions that creates the variability in the statistic calculated using bootstrapping.

Our second set of findings are shown in figure 2. Firstly, this shows that all sampling procedures add uncertainty compared to the reference procedure (which is our ground truth), and this effect increases with reduction in subsample size. Thus, the sampling procedures create variability, but (since the central tendency is accurate) not bias.

Further, figure 2 shows that, while (as shown in figure 1) the central tendency across many repetitions (of the basic procedure) gives the same dispersion estimate for *indirect_wi_rep* and *direct_wi_rep*, the former is dramatically more variable in its dispersion estimate than the latter. This then means that a single run of the basic procedure, is likely to generate an inaccurate estimate of the dispersion of a statistic if *indirect_wi_rep* is used. However, since *direct_wi_rep* exhibits the same central tendency for the dispersion of the statistic, it can be used in place of *indirect_wi_rep*, with no loss and indeed, this is exactly what is done in the main body of this paper and Lorca-Puls et al.

This is important since, as previously discussed, *indirect_wi_rep* could be considered conceptually bullet-proof, since the probability of overlap between subsamples does not reduce as the sample size gets smaller.

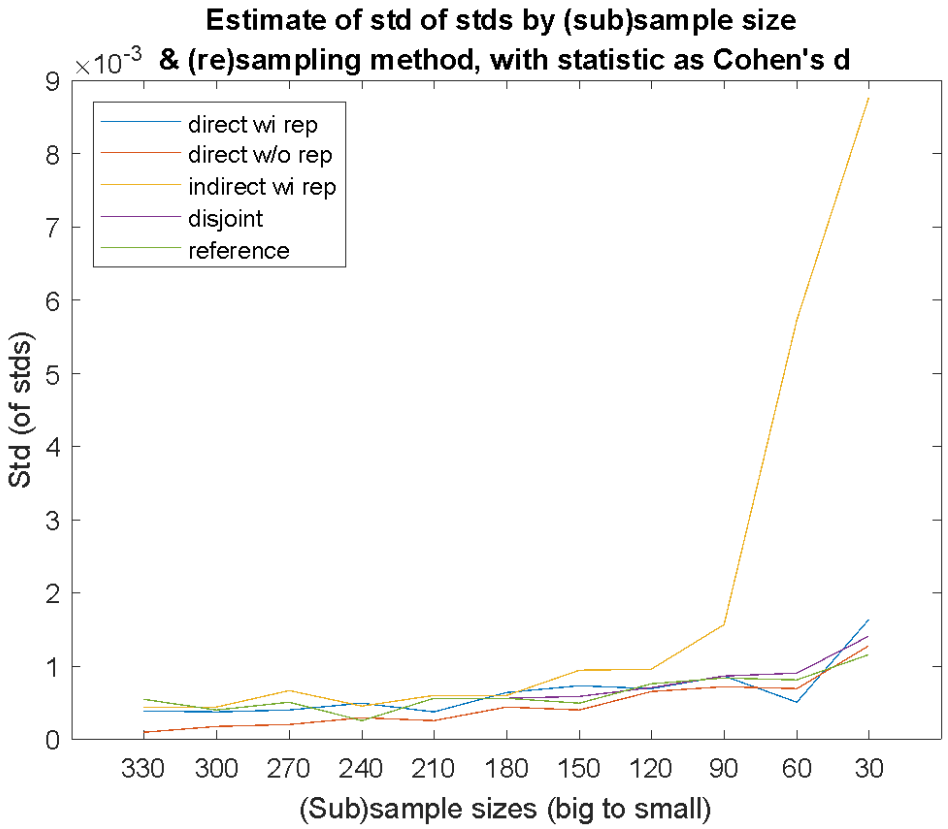

Figure 2: estimate of standard deviation of basic procedure standard deviations, when the "main statistic" is the effect size. This corresponds to a dispersion estimate of the standard deviations of basic procedure distributions. This is given for a range of sub-sample sizes. This is plotted for the four different resampling procedures, with the nature of the disjoint split procedure meaning that it cannot have subsample sizes greater than 180. The "ground truth" *reference* procedure is also shown. It is clear that *indirect_wi_rep* adds considerable variability to standard deviation estimates, as sample sizes get small.

The comparison of *reference* and *indirect_wi_rep* is also of note. The point of interest for this appendix is what happens with direct sampling with replacement (*direct_wi_rep*) as the sample size gets small, and whether there remains a non-trivial probability of the same data points appearing in multiple samples even when the sample size is small. This is because, it is the difference of this probability of overlap between direct and indirect sampling that we are interested in. Indeed, the most liberal test of our hypothesis (i.e. giving the greatest opportunity to find a disparity between direct and indirect sampling) is one in which the probability of the same data points appearing in multiple samples with direct sampling is vanishingly small. This is what the *reference* case gives us. With real numbers, there is (mathematically) a probability of zero of sampling the same number multiple times from a Gaussian, so all the samples generated under *reference* are disjoint by construction. (This is, of course, putting aside issues of maximum precision available on a particular computer, which no approach can overcome.) The fact that *reference* shows the same pattern as

direct_wi_rep in figures 1 and 2, suggests that increasing the size of the *base_sample* in order to reduce sample overlap in *direct_wi_rep* will not change the basic findings, i.e. what we demonstrate here with base samples of 360 is in fact fully general.

References

Lorca-Puls, D. L., Gajardo-Vidal, A., White, J., Seghier, M. L., Leff, A. P., Green, D. W., ... & Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. Neuropsychologia, 115, 101-111.