# Bootstrapping with Single Sweeps and Peak measures, the Issue in an Example

Bowman, H. [1,2]

(plus, Zoumpoulaki, A. [2], and Alsufyani, A. [2])

[1] School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT

and

[2] Centre for Cognitive Neuroscience & Cognitive Systems, School of Comp, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF

(Email: H.Bowman@kent.ac.uk)

## Introduction

To understand the nature of the hypothesized bias with bootstrapping with peak measures in ERP analyses, it is valuable to identify examples that demonstrate the essence of the issue. To comprehend that essence clearly it makes sense to abstract away from the complexity that is associated with real examples, but cloud understanding. This is what we do here, we present an example that is in a sense canonical, i.e. that is characteristic of the issue with bootstrap and demonstrate it in a comprehensible fashion. It is though not fully realistic; it is an idealised illustration. Separate work needs to investigate how frequently these types of situation arise, in exactly what contexts and with what severity.

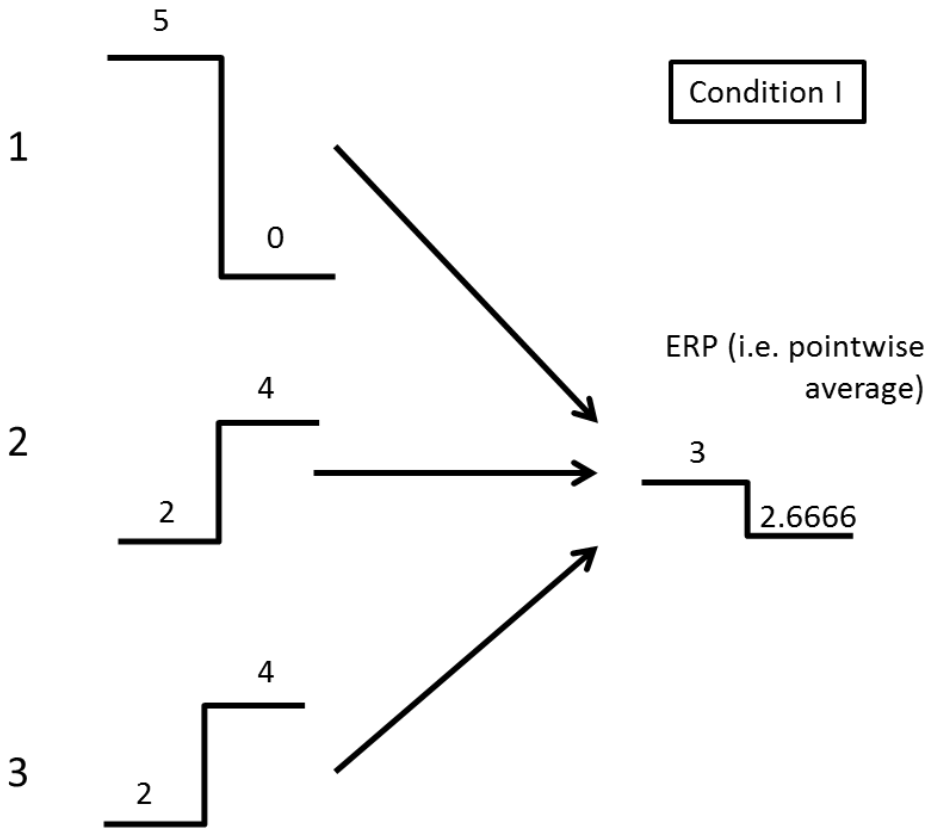
## Example: a potential false positive example

We present an inference problem in which condition I and II are in fact equal under a standard ERP contrast on the peak amplitude, but a bootstrapping test identifies a clear difference between conditions. If this introduced difference is big enough relative to the variance, bootstrapping in this case would generate a false positive.

In this example, the “true observed” ERPs have identical peak amplitudes in condition I and II. Thus, the null hypothesis is true by construction and the distribution of differences (C.I minus C.II) of peak amplitudes should be centred at zero, and also symmetric around zero.

### Condition I

This is a condition in which the standard ERP bootstrap method overestimates the peak amplitude a lot. It is a condition with three single sweeps – shown on the left, and the ERP from those three on the right.




The true peak amplitude here is 3, i.e. the amplitude of the peak of the ERP.

The set of all possible (ordered) bootstrap resamplings that can be made from three data points, here 1, 2 and 3 (for the three single sweeps), is shown in the appendix, section A.1. There are 27 such ordered resamplings.

In fact, the order of selection is unimportant. So, for example, 223 and 322 are the same, since the order in which points (or here time series) are entered into an average is irrelevant to the value it produces. So, we present the number of times each resampling arises, with order ignored, along with the amplitude of the peak in that case (i.e. taken from the surrogate ERP for each sample).

| Sample | Number of times occurs | Amplitude of peak |
|---|---|---|
| 111 | 1 | 5 |
| 112 | 3 | 4 |
| 113 | 3 | 4 |
| 122 | 3 | 3 |
| 123 | 6 | 3 |
| 133 | 3 | 3 |

| | | |
|---|---|---|
| 222 | 1 | 4 |
| 223 | 3 | 4 |
| 233 | 3 | 4 |
| 333 | 1 | 4 |

The following table summarises this, by presenting the number of times each peak amplitude arises, and the resulting probability of occurrence.

| Amplitude of peak | Frequency | Probability |
|---|---|---|
| 5 | 1 | 1/27 |
| 4 | 14 | 14/27 |
| 3 | 12 | 12/27 |
| --- | --- | --- |
| | 27 | 1 |

Then we can calculate the size of each peak weighted by its probability of occurring in the bootstrap resampling.

A = 5 x 1/27 = 0.185185

B = 4 x 14/27 = 2.0740770

C = 3 x 12/27 = 1.3333333

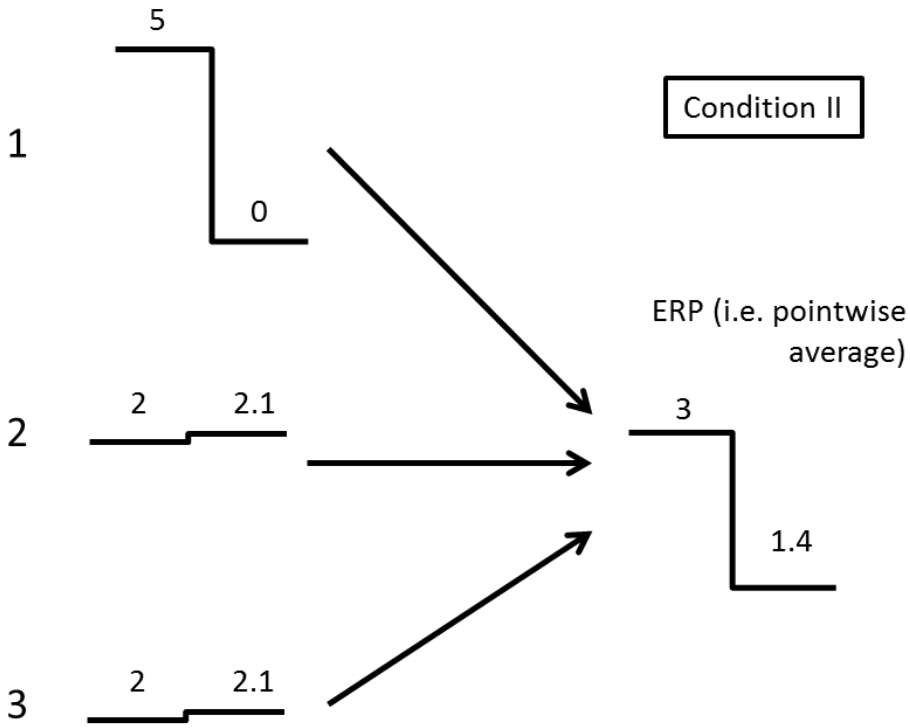
From here we can calculate the mean size of the peak across bootstrap samples, i.e. the mean of the distribution of peak amplitudes generated under bootstrapping single sweeps.

A+B+C = 3.5655183 ≠ 3

3 is the correct peak, so this overestimates the peak a lot.

## Condition II

This is a condition that overestimates the peak, but much less.




Again, the true peak amplitude is 3, i.e. of the "true-observed" ERP, and we repeat the analysis we performed for condition I.

| Sample | Number of times occurs | Amplitude of peak |
|---|---|---|
| 111 | 1 | 5 |
| 112 | 3 | 4 |
| 113 | 3 | 4 |
| 122 | 3 | 3 |
| 123 | 6 | 3 |
| 133 | 3 | 3 |
| 222 | 1 | 2.1 |
| 223 | 3 | 2.1 |
| 233 | 3 | 2.1 |
| 333 | 1 | 2.1 |

This can be summarised in the following table.

| Amplitude of peak | Probability | Peak amplitude weighted by probability |
|---|---|---|
| 5 | 1/27 | A = 5 x 1/27 = 0.1851851 |

| | | |
|---|---|---|
| 4 | 6/27 | B = 4 x 6/27 = 0.8888888 |
| 3 | 12/27 | C = 3 x 12/27 = 1.333333 |
| 2.1 | 8/27 | D = 2.1 x 8/27 = 0.62 |

Then the mean size of the peak amplitude across bootstrap samples is the following.

A+B+C+D = 3.0296 ≠ 3

So, the peak amplitude is again overestimated, but by much less than it was in condition I. As a result, under bootstrapping, condition I will often generate samples (of single sweeps) in which the peak of the surrogate ERP has a higher amplitude than it does for condition II. As a result, we can say that a bootstrapping procedure applied to single sweeps of condition I versus condition II would generate a distribution of the difference of peak amplitudes in which the majority of the probability mass is greater than zero. If this (greater than zero) mass is greater than 1-α then one would have a false positive, for a condition where the null hypothesis is true by construction, and indeed no noise has been added to the example.

## Permutation Test of Condition I vs Condition II

We now apply a permutation test to infer a significance for the difference of peak amplitudes of condition I and of condition II (actually in fact of their ERPs, in the normal way).

We do this by first creating a set of all single sweeps:

{ I.1, I.2, I.3, II.1, II.2, II.3 }

There are 20 possible ways in which this set can be split into two sets each of size three, i.e. selections of three from a set of six (notice, once 3 are selected for surrogate condition I, the items in condition II are given automatically as the remainder). The appendix (section A.2) contains an enumeration of all these. Each of these "splittings" is a permutation of the single sweeps. If you work this through, i.e. for each permutation generate an ERP for surrogate Condition I, determine its peak amplitude, and subtract it from the same for Condition II, one finds the following frequencies with which differences occur:

| peakAmp(surrogate(I)) minus peakAmp(surrogate(II)) | frequency |
|---|---|
| -1.2666666 | 2 |
| -.63333333 | 2 |
| 0 | 12 |
| .633333333 | 2 |
| 1.26666666 | 2 |

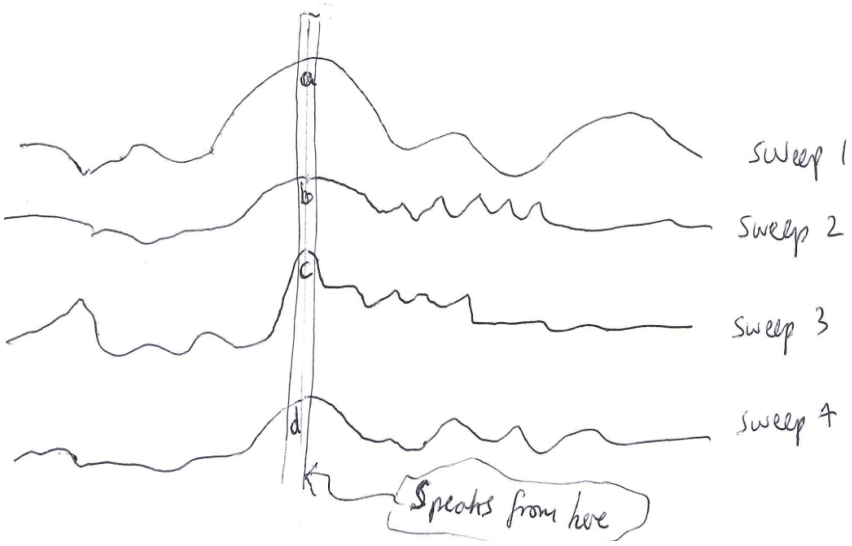
Now, this distribution is centred at zero and symmetric. Since the true observed difference of peak amplitudes is zero, the P-value the permutation test would generate is 0.5, which is what it should be, if there is no difference between the two conditions.

## Heart of the Issue

A key point to keep in mind is that the (original) true-observed peak is calculated (by taking the mean) from the set of points across sweeps at the time position where that peak occurs. If we take *those same* points and bootstrap from them, and use a "safe" statistic, such as mean, there will be no bias with bootstrapping. The critical issue is that these points are sampled from evenly, i.e. such that each point has an equal probability of entering the bootstrap sample on each selection.

In this context, the following situations attempt to clarify exactly what the issue is with bootstrapping. Specifically, the first two points we present give examples of bootstrap resampling that is not biased[1].

1) Bootstrapping EEG single sweeps to infer a peak amplitude (i.e. the standard approach) works fine when all single sweeps have their peak *at the same point* (we call this the "time-invariant peak" case), e.g.




The critical issue is that on every bootstrap resampling a set[2] of points contribute to the surrogate ERP-peak amplitude, and that set is always a subset of the set $S_{peaks}$ containing the peak amplitudes of all single sweeps. So, effectively this case becomes analogous to taking the peak amplitudes of every single sweep, creating a set of all these peak amplitudes, and bootstrapping from that set with a mean statistic[3] (i.e. taking the mean of each bootstrapped sample and placing that in a distribution). Furthermore, this set $S_{peaks}$ is exactly the set that is averaged to generate the (original) true observed peak amplitude. Importantly, bootstrapping from a fixed set with a mean statistic is an unbiased procedure; see Zoumpoulaki, Alsufyani and Bowman, 2014.

[1] Note, this is not an assessment of these two situations effectiveness in the sense of statistical power, which may be very low.
[2] Strictly each resampling gives a bag\ family of points, not a set, but consistent with usage in this literature, and to simplify presentation, we talk in terms of the set that underlies the bag.
[3] Taking the mean here results from the generation of the ERP.

2) Another alternative to the (standard) bootstrapping sweeps method that would not be biased would be to pick the time point of the peak in the original (observed) ERP, and then calculate the mean on each surrogate ERP *at that same time point*. This is not biased, since it would also reduce to bootstrapping from a fixed set of numbers (the values from the single sweeps at the point of the original (true-observed) ERP peak, and which contribute to its amplitude) with a mean statistic.

3) The standard ERP bootstrapping of single sweeps with peak measures goes wrong when the properties underlying the previous two points in this list do not obtain. Our potential false positive example is just such a case. And the critical point is that there is typically temporal jitter between single sweeps, so peaks will arise at different time points in the single sweeps. When this is combined with pointwise averaging associated with generation of the surrogate ERP from a bootstrapped sample of single sweeps, the problem arises.

Characteristic of this situation is that the single sweeps containing the largest amplitude peaks are not selected in the resample[4]. Often these are the amplitudes that are driving the placement of the peak in the (original) true-observed ERP. As a result, the peak in the surrogate ERP arises at a different time point to where it appears in the original (true-observed) ERP. The consequence of which is that a set of values contribute to the (point-wise) mean (that is placed in the resulting statistical distribution) that could not contribute to the (point-wise) mean if the peak occurred where the original (true-observed) peak occurred in time. The key issue is that the peak in the surrogate ERP can occur at a different time point depending upon the single sweeps that make it into the bootstrap sample

This is what we see in our example – if single sweep one does not get selected, the peak in the surrogate ERP will be at the second time point, rather than the first. As a result, a different set of values will contribute to the mean (since ERPs are point-wise means) entered into the statistic distribution[5].

One way to illustrate this is to see how this third situation is different to those in the previous two situations, where the bias does not arise. The central issue is that in the previous two situations, effectively, a *fixed* set of numbers is being sample from with a mean statistic and those are the numbers in single sweeps at the time point that the peak occurs at in the original (true-observed) ERP, and which thus contribute to the (true-observed) value. Furthermore, those numbers are being sampled from evenly. The standard ERP bootstrapping of single sweeps with peak measures does not always do this. In particular, the set of numbers across the single sweeps at the position of the original (true-observed) ERP peak are not evenly sampled from. Characteristically, samples dominated by the low numbers in this set do not "win" against peaks at other positions, and thus

[4] This is a sense to which the problem in the ERP case is like that in the classic max bootstrapping case discussed in the mathematical statistics literature, and, illustrated in the first two examples in Zoumpoulaki, Alsufyani and Bowman, 2015, with results shown in figure 1 of that paper. The difference being that with ERPs, the problem is about single sweeps containing big peak amplitudes being missed out of the sample, and thus not contributing to the surrogate ERP, rather than big individual values being missed out of bootstrapped samples. So, the line of explanation is more involved in the ERP case but it is inherently similar.

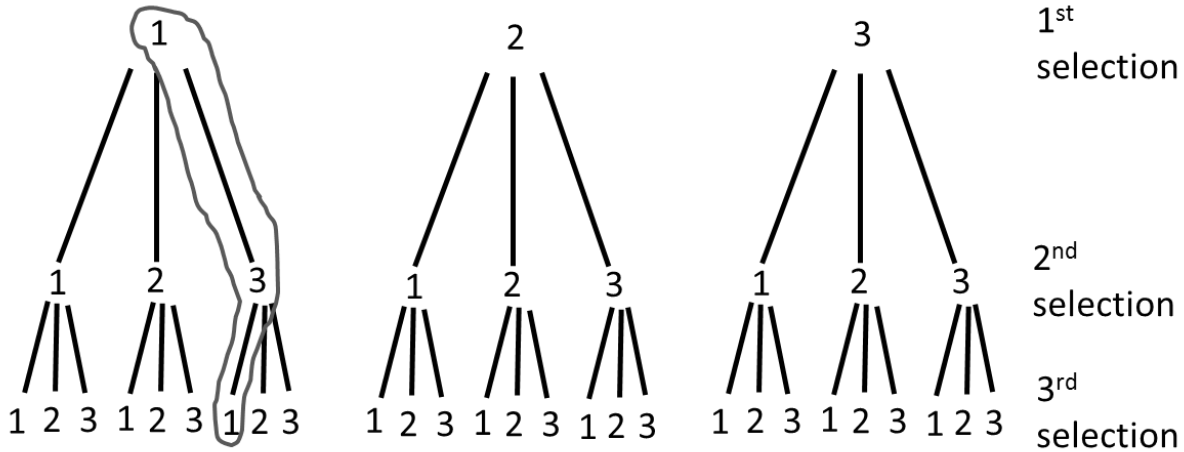
[5] This implies that there are dependencies in the resampling process, i.e. dependent upon what single sweeps make it into the bootstrap sample, a different set of numbers could contribute to the statistic (the mean) entering the statistic distribution. Such dependencies induce biases in the resulting statistical inference.

contribute less to means entered into the statistic distribution. That is why in the example, we see bootstrapping *overestimating* the true peak[6].

## Appendix

### A.1 Bootstrap Resamplings

We enumerate here the set of possible sequences of three selections when sampling with replacement.




There are 27 possible passes that one can make through these trees, giving us the following set of ordered resamplings with replacement. **131** is the sequence shown in the above figure.

| | | |
|---|---|---|
| 111 | 211 | 311 |
| 112 | 212 | 312 |
| 113 | 213 | 313 |
| 121 | 221 | 321 |
| 122 | 222 | 322 |
| 123 | 223 | 323 |
| **131** | 231 | 331 |
| 132 | 232 | 332 |
| 133 | 233 | 333 |

[6] The amount of this overestimation will vary as a function of how big the signal is in a particular condition, with, I currently think, smaller overestimation as the signal becomes bigger relative to the noise. This is why overestimating the peak in this way can lead to an underestimation of the difference of peak amplitudes measure.

## A.2 Permutation Resamplings

Now, we move on to enumerating all the possible permutation resamplings. The following contains all these, with the difference that would arise from each permutation.

| Surrogate Cond I | Surrogate Cond II | peakAmp(surrogate(I)) minus peakAmp(surrogate(II)) | Result |
|---|---|---|---|
| I.1, I.2, I.3 | II.1, II.2, II.3 | 3-3 | 0 |
| II.1, I.2, I.3 | I.1, II.2, II.3 | 3-3 | 0 |
| II.2, I.2, I.3 | II.1, I.1, II.3 | 10.1/3 – 12/3 | -0.633333 |
| II.3, I.2, I.3 | II.1, II.2, I.1 | 10.1/3 – 12/3 | -0.633333 |
| I.1,II.1, I.3 | I.2, II.2, II.3 | 12/3 – 8.2/3 | 1.2666666 |
| I.1,II.2, I.3 | II.1, I.2, II.3 | 9/3 – 9/3 | 0 |
| I.1,II.3, I.3 | II.1, II.2, I.2 | 9/3 – 9/3 | 0 |
| I.1, I.2, II.1 | I.3, II.2, II.3 | 12/3 – 8.2/3 | 1.2666666 |
| I.1, I.2, II.2 | II.1, I.3, II.3 | 9/3 – 9/3 | 0 |
| I.1, I.2, II.3 | II.1, II.2, I.3 | 9/3 – 9/3 | 0 |
| II.1,II.2, I.3 | I.1, I.2, II.3 | 9/3 – 9/3 | 0 |
| II.1,II.3, I.3 | I.1, II.2, I.2 | 9/3 – 9/3 | 0 |
| II.2,II.3, I.3 | II.1, I.1, I.2 | 8.2/3 - 12/3 | -1.2666666 |
| I.1,II.1, II.2 | I.2, I.3, II.3 | 12/3 – 10.1/3 | 0.633333 |
| I.1,II.1, II.3 | I.2, II.2, I.3 | 12/3 – 10.1/3 | 0.633333 |
| I.1,II.2, II.3 | II.1, I.2, I.3 | 9/3 – 9/3 | 0 |
| II.1,I.2, II.2 | I.1, I.3, II.3 | 9/3 – 9/3 | 0 |
| II.1,I.2, II.3 | I.1, II.2, I.3 | 9/3 – 9/3 | 0 |
| II.2,I.2, II.3 | II.1, I.1, I.2 | 8.2/3 - 12/3 | -1.2666666 |
| II.1, II.2, II.3 | I.1, I.2, I.3 | 3-3 | 0 |