

With great power comes greater responsibility

Last compiled: Tuesday 28th March, 2023 at 19:33

1 Introduction

A standard problem in psycholinguistics **not really limited to psychology...processes unfolding in time are pretty ubiquitous throughout science**, and the cognitive sciences in general, is that of statistically analyzing a process unfolding in time.

Particularly in the case of comparing a process in time between experimental groups, while there are many techniques for demonstrating *that* a difference exists, few offer any insight into *when* it exists. Testing for temporal differences is complicated by the fact that there are often an arbitrarily large number of time points that could be compared, depending on the sampling density, and where significance tests at nearby time points have the potential to be highly correlated. This is essentially a problem of multiple comparisons.

Various approaches have been proposed for addressing this issue. The most rudimentary of these involve binary tests for identifying the existence of any difference, such as the area under the curve (AUC). More sophisticated techniques involve the use of cluster-based permutation testing (?), whereby test statistics are computed at each observed time point and with adjacent significant tests being combined into clusters. This controls the family-wise error rate (FWER) by reducing adjacent test statistics into a single cluster, thereby reducing the number of total tests. More recently, ? introduced a modified Bonferroni correction to a series of test statistics, using estimates of autocorrelation between statistics to make the appropriate adjustments to the significance levels to control FWER. This approach, which they named bootstrapped differences in time series, was introduced in the R package, **bdots** (?).

The modified Bonferroni correction used by **bdots** relies on the construction of estimated distributions of time series via bootstrapping for each experimental group. These distributions, in turn, are used to construct test statistics at each observed time point. A closer look at the original iteration presents concerns as it involves quite restrictive assumptions on the data that are unlikely to be met in many, if not most situations. This includes data typically collected in the Visual World Paradigm **should i have a paragraph describing**

VWP? not sure where to squeeze in, the context in which the underlying methodology was first proposed. Specifically, the original implementation assumes a homogeneous mean structure within each experimental group, with no between-subject variability to be accounted for. Empirical data collected in a variety of (VWP) contexts can be used to demonstrate that this assumption is unlikely to be true, the consequence of which is a type I error rate that is unacceptably high.

Here, we present two alternative methods that accommodate flexibility in the assumptions made by the original bootstrapped differences in time series algorithm. First, we propose a modified bootstrapping procedure that adequately accounts for observed between-subject variability while retaining the FWER adjustment that addresses the autocorrelated test statistics. In addition, we offer a permutation test for identifying temporal differences between groups, borrowing from the insight of the original bdots in that it also captures within-subject variability as demonstrated in the standard errors in the model fits.

We begin with a mathematical description of the problem along with the proposed alternatives to the original bootstrapping algorithm. This is followed by a simulation estimating the FWER across a number of experimental conditions. Finally, we consider two separate simulations for the assessment of power among the competing methods. This includes one simulation with a simple piecewise-linear function relating power to effect sizes; the second simulation interrogates power when the magnitudes of between-subject and within-subject variability differ. [is this too much detail here?](#)

2 Methods

In each of the methods to be described, we begin with the observation of y_{it} for subjects $i = 1, \dots, n$ over times $t = 1, \dots, T$. Typically, these subjects fall into different groups $g = 1, \dots, G$, with each group containing n_g subjects, and it is these groups we wish to compare in time. We further assume that the empirically observed data follows from a parametric function f with associated error:

$$y_{it} = f(t|\theta_i) + \epsilon_{it} \tag{1}$$

where θ_i is the subject-specific parameterization of f and

$$\epsilon_{it} = \phi\epsilon_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma). \tag{2}$$

Under this paradigm, the errors could be IID normal (with $\phi = 0$) or have an AR(1) structure, with $0 < \phi < 1$. It is generally assumed that the observed data across subjects make up a distribution for each group, either a multivariate distribution of the parameters θ_i or a distribution of resulting curves, $f(t|\theta_i)$.

Ultimately, it will be from a distribution of curves that we determine the temporal characteristics of each group, with the differences in these characteristics specifically being what we are interested in identifying.

With the general notation addressed, we now move to the particulars of each of the methods considered.

2.1 Homogeneous bootstrap

The original bootstrapping algorithm presented in ? follows what we will call the *homogeneous means* assumption. Accordingly, we will call this bootstrapping algorithm the *homogeneous bootstrap*. Under the homogeneous means assumption, it is still assumed that observed data for each subject retains the mean structure given in Equation 1, but with the additional assumption that $\theta_i = \theta_j$ for all subjects i, j within the same group. In other words, there is assumed to be no variability in the mean structure between subjects within the same group. This is evidenced in the original bootstrapping differences in time series algorithm:

1. For each subject, fit a nonlinear regression model to obtain $\hat{\theta}_i$. ? recommended specifying an AR(1) autocorrelation structure for model errors. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a multivariate normal distribution with mean for subject i , $\hat{\theta}_i$ corresponding to the point estimate and standard deviations corresponding to the standard errors, s_i .
2. Using the approximate sampling distributions in (1.), randomly draw one bootstrap estimate for each of the model parameters on every subject

$$\hat{\theta}_i^{(b)} \sim N(\hat{\theta}_i, s_i^2) \quad (3)$$

3. Once a bootstrap estimate has been collected for each parameter and for every subject, for each parameter, find the mean of the bootstrap estimates across n_g individuals for the b th bootstrap in group g ,

$$\theta_g^{(b)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \hat{\theta}_i^{(b)} \quad (4)$$

4. Use the mean parameter estimates to determine a bootstrapped population level curve, which provides the average population response at each time point, $f(t|\theta_g^{(b)})$.
5. Perform steps (2)-(4) 1000 times to obtain estimates of the population curves. Use these to create estimates of the mean response and standard deviation at each of the time points. For each group

$g = 1, \dots, G$, this gives

$$\bar{p}_{gt} = \frac{1}{B} \sum_{b=1}^B f(t|\theta_g^{(b)}), \quad s_{gt}^2 = \frac{1}{B-1} \sum_{b=1}^B f(t|\theta_g^{(b)}) - \bar{p}_{gt}. \quad (5)$$

Population means and standard deviations at each time point for each of the groups were used to construct a series of (correlated) test statistics, where the family wise error rate was controlled by using the modified Bonferonni correction introduced in ? to test for significance. **This modified correction seems kind of important (among other things, hetboot uses it too), I think you should say more about it**

2.2 Heterogeneous Bootstrap

Typically, subjects within a group demonstrate considerable variability in their mean parameter estimates. In this case, we should avoid the presumption that $\theta_i = \theta_j$ as accounting for between-subject variability within a group will be critical for obtaining a reasonable distribution of the population curves. More likely, we may assume that the distribution of parameters for subjects $i = 1, \dots, n_g$ in group $g = 1, \dots, G$ follows the distribution

$$\theta_i \sim N(\mu_g, V_g), \quad (6)$$

where μ_g and V_g are the group-specific mean and variance values, respectively. Similar to what was presented in the homogeneous bootstrap algorithm, we can further account for uncertainty in our estimation of θ_i by $\hat{\theta}_i$ by treating the standard errors derived when fitting the observed data to the mean structure suggested in Equation 1 as estimates of their standard deviations. This gives us a multivariate normal distribution for each subject's estimated parameter,

$$\hat{\theta}_i \sim N(\theta_i, s_i^2). \quad (7)$$

As our goal remains as being able to obtain reasonable estimates of the population curves for each group, it is necessary to estimate both the observed within-subject variability found in each of the $\{s_i^2\}$ terms, *as well as* the between-subject variability present in V_g . For example, let θ_{ib}^* represent a bootstrapped sample for subject i in bootstrap $b = 1, \dots, B$, where

$$\theta_{ib}^* \sim N(\hat{\theta}_i, s_i^2), \quad (8)$$

as was done in Step (2.) of the homogeneous bootstrapping algorithm. If we were to sample *without*

replacement, we would obtain a homogeneous mean value from the b th bootstrap for group g , $\theta_{bg}^{(hom)}$, where

$$\theta_{bg}^{(hom)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \theta_{ib}^*, \quad \theta_{bg}^{(hom)} \sim N \left(\mu_g, \frac{1}{n_g^2} \sum_{i=1}^{n_g} s_i^2 \right). \quad (9)$$

Such an estimate captures the totality of the within-subject variability with each draw but fails to account for the variability in the group overall. For this reason, we sample the subjects *with* replacement, creating the heterogeneous bootstrap mean $\theta_{bg}^{(het)}$, where again each θ_{ib}^* follows the distribution in Equation 8, but the heterogeneous bootstrapped group mean now follows [indicated in hom bootstrap that it was without replacement via the enumerators on the summation, not sure how to precisely indicate that here](#)

$$\theta_{bg}^{(het)} \sim N \left(\mu_g, \frac{1}{n_g} V_g + \frac{1}{n_g^2} \sum s_i^2 \right). \quad (10)$$

The estimated mean value remains unchanged, but the variability is now fully accounted for. We therefore present a modified version of the bootstrap which we call the *heterogeneous bootstrap*, making the following changes to the original: [I'm not sure what math to include here, how else indicate im sampling without replacement](#)

1. In step (1.), the specification of AR(1) structure is *optional* and can be modified with arguments to functions in `bdots`. Our simulations show that while failing to include it slightly inflates the type I error in the v2 bootstrap when the data truly is autocorrelated, specifying an AR(1) structure can lead to overly conservative estimates when it is not.
2. In step (2.), we sample subjects *with replacement* and then for each drawn subject, randomly draw one bootstrap estimate for each of their model parameters based on the mean and standard errors derived from the `gnls` estimate.

Just as with the homogeneous bootstrap, these bootstrap estimates are used to create test statistics T_t at each time point, written

$$T_t^{(b)} = \frac{(\bar{p}_{1t} - \bar{p}_{2t})}{\sqrt{s_{1t}^2 + s_{2t}^2}}, \quad (11)$$

where \bar{p}_{gt} and s_{gt}^2 are mean and standard deviation estimates at each time point for groups 1 and 2, respectively. Finally, just as in Oleson 2017, one can use the autocorrelation of the $T_t^{(b)}$ statistics to create a modified α for controlling the FWER.

2.3 Permutation Testing

In addition to the heterogeneous bootstrap, we also introduce a permutation method for hypothesis testing. The permutation method proposed is analogous to a traditional permutation method, but with an added step mirroring that of the previous in capturing the within-subject variability. For a specified FWER of α , the proposed permutation algorithm is as follows:

[not sure what additional math to add here](#)

1. For each subject, fit the nonlinear function with *optional* AR(1) autocorrelation structure for model errors. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a normal distribution with mean corresponding to the point estimate and standard deviation corresponding to the standard error
2. Using the mean parameter estimates derived in (1.), find each subject's corresponding fixation curve. Within each group, use these to derive the mean and standard deviations of the population level curves at each time point, denoted \bar{p}_{jt} and s_{jt}^2 for $j = 1, 2$. Use these values to compute a test statistic T_t at each time point,

$$T_t^{(p)} = \frac{|\bar{p}_{1t} - \bar{p}_{2t}|}{\sqrt{s_{1t}^2 + s_{2t}^2}}. \quad (12)$$

This will be our observed test statistic.

3. Repeat (2) P additional times, each time shuffling the group membership between subjects. This time, when constructing each subject's corresponding fixation curve, draw a new set of parameter estimates using the distribution found in (1). Recalculate the test statistics $T_t^{(p)}$, retaining the maximum value from each permutation. This collection of P statistics will serve as our null distribution which we denote \tilde{T} . Let \tilde{T}_α be the $1 - \alpha$ quantile of \tilde{T}
4. Compare each of the observed $T_t^{(p)}$ with \tilde{T}_α . Areas where $T_t^{(p)} > \tilde{T}_\alpha$ are designated significant.

2.4 Paired data

Briefly, we attend to the issue of paired data for each of the discussed methods, as this is critical for a proper assessment of both FWER and power. Specifically, in a paired setting we note that our interest is in determining the distribution of differences rather than of the respective groups.

For the homogeneous bootstrap, this is done by default: as each subject is sampled without replacement, we can be sure that each bootstrap estimate between groups contains the same subjects. For the heterogeneous bootstrap, this is done by ensuring that the same subjects sampled in each bootstrap for one group

is matched identically with sampling subjects in the other. Put differently, this states that we begin by determining which subjects will be included in each bootstrap *for both groups at the same time*. Lastly, in the case of permutation testing, paired data is addressed by ensuring that each permuted group contains one observation for each subjects, so that each subject in one permuted group has its corresponding observation in the other. [am i saying this weird?](#)

3 FWER Simulations

We now go about comparing the type I error rate of the three methods just described. In doing so, we will consider several conditions under which the observed subject data may have been generated or fit. This includes generating data with both a homogeneous and heterogeneous means assumption, generating data with and without autocorrelated errors, and fitting data with and without an AR(1) assumption.

In addressing performance under paired and unpaired conditions, we have included two distinct but reasonable instances in which the data may be paired, which we will elaborate further on shortly. However, as paired data becomes difficult to define under the homogeneous means assumption (in which all subjects are, in a sense, “paired”), we will omit these settings from our final simulations. In all, this gives us sixteen different arrangements which we will examine for their family wise error rates using each of the three methods previously described.

We turn now to a detailed description of how the data was generated.

3.1 Data Generation

Data was generated according to Equation 1, with the parametric function $f(t|\theta)$ belonging to the family of four-parameter logistic curves defined:

$$f(t|\theta) = \frac{p - b}{1 + \exp\left(\frac{4s}{p-b}(x - t)\right)} + b \quad (13)$$

where $\theta = (p, b, s, x)$, the peak, baseline, slope, and crossover parameters, respectively.

We further assume that each group drew subject-specific parameters from a normal distribution, with subject $i = 1, \dots, N$ in group $g = 1, \dots, G$ following the distribution in Equation 6. These parameters are then used to simulate empirical data according the the mean and error structures for each simulation setting, which we detail next.

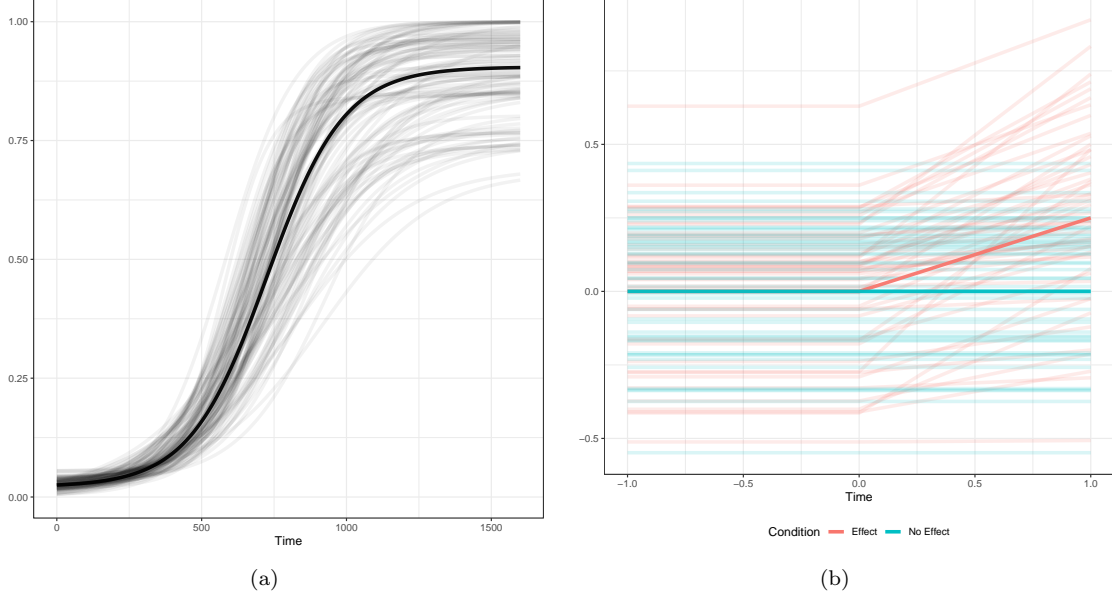


Figure 1: Distributions of various group with the mean curve in bold (a.) 50 samples from the generating distribution of the four-parameter logistic in Equation 13 used for testing FWER. (b.) 50 samples from the generating distributions of each group in Equation 16. The legend makes size weird, might just explain what they are here. Also need to see if I can change size of the mean lines

Mean Structure In all of the simulations presented, the distribution used in Equation 6 was empirically determined from data on normal hearing subjects in the VWP (?). Parameters used were those fit to fixations on the Target, following the functional form of Equation 13. We will call this the heterogeneous means assumption.

Under the assumption of between-subject homogeneity, we set $\theta_i = \theta_j$ for all subjects i, j , assuring that each of the subjects' observations is derived from the same mean structure, differing only in their observed error. We will call this the homogeneous means assumption.

Error Structure The error structure is of the form

$$e_{it} = \phi e_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma) \quad (14)$$

where the w_{it} are iid with $\sigma = 0.025$. ϕ corresponds to an autocorrelation parameter and is set to $\phi = 0.8$ when the generated data is to be autocorrelated and set to $\phi = 0$ when we assume the errors are all independent and identically distributed.

Paired Data As we previously noted, paired data is only a sensible condition under the assumption of heterogeneous means, and we limit our consideration to that case. Further, we observe that in the

construction of paired data, there are two methods that seem reasonable, and we employ both of them here. In considering the construction of paired data for subject i , the first method proceeds as follows: we begin by drawing parameters θ_i from Equation 6. Denote this θ_{i1} to indicate that this is the parameter estimate for subject i in group 1. We then simulate observed data according to Equation 1. To create the paired data, we then set $\theta_{i2} = \theta_{i1}$ and again simulate observed data according to Equation 1. Critically, under this first method, the generating parameters between groups are *identical*, with the only differences between the simulated data being that contributed by the error term.

In the second method of obtaining paired data, we proceed as in the first, except now letting

$$\theta_{i2} = \theta_{i1} + N(0, 0.05 \cdot V_g). \quad (15)$$

This adds a small amount of random noise between paired parameters, simulating the degree of variability that may normally be found between conditions, even when there is no true effect. To avoid potential confusion, the results for each of these will be presented separately. [I haven't decided how to address these yet in columns.](#)

Each set of conditions generates two groups, with $n = 25$ subjects in each group, with time points $t = 0, 4, 8, \dots, 1600$ in each trial and with 100 simulated trials for each subject. Columns in the tables indicate homogeneity of means assumption, whether or not an AR(1) error structure was used in constructing the data, and if autocorrelation was specified in the fitting function. The last conditions help assess the impact of correctly or incorrectly identifying the type of error when conducting an analysis in `bdots`. Each simulation was conducted 1000 times to determine the family wise error rate.

3.2 Results

We consider the efficacy the methods under each of the simulation settings with an analysis of the family wise error rate (FWER) and the median per-comparison error rate. The first of these details the proportion of simulations under each condition that marked *at least* one time point as being significantly different between the two groups. This is critical in understanding each method's ability to correct adjust for the multiple testing problem associated with testing each of the observed time points. These are presented in Table 1 and Table 2 for unpaired and paired data, respectively.

Complementing the FWER estimate is an estimate of the median per-comparison rate. For each time point across each of the simulations, we computed the proportion of times in which that time was determined significant. The median of these values across all time points is what is considered. This metric gives a sense of magnitude to the binary FWER; for example, a situation in which there was a high FWER and low

median per-comparison rate would indicate that the type I error within a particular time series would be sporadic and impact limited regions. Large median per-comparison rates indicate that large swaths of a time series frequently sustain type I errors. The median per-comparison rates for unpaired and paired simulations are presented in Table 4 and Table 5.

3.2.1 FWER

There are a few things of immediate note when considering the results of Table 1. First, we see from the first two settings of the unpaired simulations that the type I error rates for the homogenous bootstrap are consistent with those presented in ?, confirming the importance of specifying the existence of autocorrelation in the `bdots` fitting function when autocorrelated error is present. By contrast, this is far less of a concern when using the heterogeneous bootstrap or permutation testing, both of which maintain a FWER near the nominal alpha, regardless of whether or not the error structure was correctly identified. This continues to be true under the homogenous mean assumption when the true error structure is not autocorrelated. Interestingly, the performance of the homogeneous bootstrap falters here despite theoretical consistency with the simulation settings. [I am rerunning this condition again now to make sure there wasn't an error.](#)

The most striking results of this, however, appear when the data generation assumes a heterogeneous mean structure. While both the heterogeneous bootstrap and the permutation test maintain a FWER near the nominal alpha, the homogeneous bootstrap fails entirely, with a FWER > 0.9 in all cases.

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
No	Yes	Yes	0.09	0.00	0.06
No	Yes	No	0.84	0.06	0.14
No	No	Yes	0.12	0.01	0.08
No	No	No	0.14	0.00	0.05
Yes	Yes	Yes	0.94	0.05	0.05
Yes	Yes	No	0.99	0.07	0.07
Yes	No	Yes	1.00	0.08	0.05
Yes	No	No	0.99	0.05	0.04

Table 1: FWER for empirical parameters (unpaired)

Paired data is given in Table 2. Matching the conclusions drawn from Table 1, we only note here that both the permutation test and heterogeneous bootstraps maintain a valid FWER under the assumption of

paired data.

Het.	AR(1)	AR(1)	Hom.	Het.	Perm.
Means	Error	Specified	Boot	Boot	
Yes	Yes	Yes	0.10	0.00	0.12
Yes	Yes	No	0.75	0.06	0.12
Yes	No	Yes	0.12	0.00	0.11
Yes	No	No	0.11	0.01	0.13

Table 2: FWER for empirical parameters (paired setting 1)

Het.	AR(1)	AR(1)	Hom.	Het.	Perm.
Means	Error	Specified	Boot	Boot	
Yes	Yes	Yes	0.48	0.04	0.10
Yes	Yes	No	0.93	0.07	0.12
Yes	No	Yes	0.81	0.04	0.08
Yes	No	No	0.81	0.07	0.09

Table 3: FWER for empirical parameters (paired settings 2)

3.2.2 Median per comparison error rate

We next consider the median comparison rate, which offers some insight into the FWER. In particular, consider the situation in which in Table 4, in the fourth row we see a median per-comparison error rate of 0.00 for the homogeneous bootstrap, despite Table 1 indicating a FWER of 0.15. This is a consequence of the majority of the type I errors occurring in a relatively limited region. In contrast, the median per-comparison error rate of the homogeneous bootstrap under the assumption of heterogeneity suggests that the type I errors are widespread and not limited to any particular area.

It is also worth commenting on the permutation test median per-comparison error rate in Table 4; combined with a FWER near the nominal 0.05, these values suggest that errors are likely distributed across the entire range rather than limited to a small area (which would result in a MPCR of 0).

[I confirmed this by looking both at this histograms and inspecting the data manually](#)

Het.	AR(1)	AR(1)	Hom.	Het.	Perm.
Means	Error	Specified	Boot	Boot	
No	Yes	Yes	0.02	0.00	0.01
No	Yes	No	0.31	0.01	0.02
No	No	Yes	0.01	0.00	0.01
No	No	No	0.00	0.00	0.01
Yes	Yes	Yes	0.59	0.01	0.01
Yes	Yes	No	0.83	0.02	0.01
Yes	No	Yes	0.84	0.02	0.01
Yes	No	No	0.82	0.01	0.01

Table 4: median per comparison error rate (unpaired)

Het.	AR(1)	AR(1)	Hom.	Het.	Perm.
Means	Error	Specified	Boot	Boot	
Yes	Yes	Yes	0.03	0.00	0.02
Yes	Yes	No	0.27	0.01	0.02
Yes	No	Yes	0.01	0.00	0.02
Yes	No	No	0.01	0.00	0.02

Table 5: median per comparison error rate (paired 1)

Het.	AR(1)	AR(1)	Hom.	Het.	Perm.
Means	Error	Specified	Boot	Boot	
Yes	Yes	Yes	0.14	0.01	0.02
Yes	Yes	No	0.46	0.01	0.03
Yes	No	Yes	0.44	0.01	0.01
Yes	No	No	0.41	0.02	0.02

Table 6: median per comparison error rate (paired 2)

3.3 Discussion

Table 1 and Table 2 demonstrate that under a variety of settings, both the heterogeneous bootstrap and permutation offer a FWER near the nominal alpha, making their performance similar to that of the homo-

geneous bootstrap in the best of cases. This includes less restrictive assumptions in which they continue to perform well while the homogeneous bootstrap maintains an unacceptably high FWER.

Transition sentence to power simulations

4 Power Simulations

To determine power, two experimental groups were simulated with mean structures of the following form:

$$y = \begin{cases} b & x < 0 \\ mx + b & x \geq 0 \end{cases} \quad (16)$$

The first simulated group was “No Effect”, with intercept and slope parameters normally distributed and standard deviation $\sigma = 0.05$. The second group, the “Effect” group, was similarly distributed, but with the slope parameter having mean value of $\mu = 0.25$. The error structure was identical to that in the FWER simulations, with both an AR(1) error structure and independent noise included. 100 simulations were conducted for each scenario.

We limited consideration to three possible scenarios: first, we assumed the conditions presented in Oleson 2017, assuming homogeneity between subject parameters and an AR(1) error structure, with the model fitting performed assuming autocorrelated errors. For the remaining scenarios, we assumed heterogeneity in the distribution of subject parameters, simulated with and without an AR(1) error structure. In both of these last two scenarios, we elected to *not* fit the model assuming autocorrelated errors. This was for two reasons: first, simulations exploring the type I error rate suggested that models fit with the autocorrelation assumption tended to be conservative. Second, and given the results of the first, this makes setting the assumption of autocorrelation to FALSE in `bdots` seem like a sensible default, and as such, it would be of interest to see how the model performs in cases in which there is autocorrelated error that is not accounted for.

For each subject, parameters for their mean structure given in Equation 16 were drawn according to their group membership and fit using `bdots` on the interval (-1,1). Time windows in which the groups differed were identified using each the homogenous bootstrap, heterogeneous bootstrap, and permutation testing. By including the interval (-1,0) in which the null hypothesis was true, we are able to mitigate the effects of over-zealous methods in determining power, and we present the results in the following way: any tests in which a difference was detected in (-1,0) was marked as having a type I error, and the proportion of simulations in which this occurred for each method is reported as the FWER in the column labeled α . The next column,

β , is the type II error rate, indicating the proportion of trials in which no differences were identified over the entire region. The last Greek-letter column is $1 - \beta - \alpha$, a modified power statistic indicating the proportion of tests in which a difference was correctly identified. The remaining columns relate to this modified power columns, giving a partial summary of the earliest onset of detection. As a true difference occurs on the interval $t > 0$, smaller values indicate greater power in detecting differences. Finally, a plot giving the power at each time point is given in Figure 2. This plot represents the true power, though note that it does not take into account the rate at which these regions were identified in conjunction with a type I error rate.

4.1 Results

[tables up until this point had *actually* been run with incorrect slope (0.025 instead of 0.25). Presented here now (3/5) are the correct tables. None of the results changed materially, though the type II error is much lower (given the corrected larger effect). Additionally, the current results only include 200 simulations. For completeness, I am running the full suite of simulations with both $m = 0.025$ and $m = 0.25$, though the only difference, again, will be in type II error and onset of detection]

The results of the power simulation are presented in Table 6. We begin by considering the case in which we assumed a homogeneous mean structure with autocorrelated errors (the first three rows), matching the conditions in which the homogeneous bootstrap was first presented. Notably, we find that the permutation method demonstrates the greater power, with the median onset time just under that of the homogeneous bootstrap. This is at the expense of a larger FWER, though still below the nominal level. Alternatively, the heterogeneous bootstrap maintains a similar FWER as the homogeneous bootstrap at the cost of power.

The remaining settings tell a similar story with the exception of the homogeneous bootstrap which continues to demonstrate unacceptable FWER under the heterogeneous means assumption. We also seem some effect of not correctly specifying an AR(1) error structure when comparing the last two settings, as the failure to specify resulted in a slightly higher FWER and lower power, although not significantly (see Appendix for tables and figures of remaining situations).

Method	Heterogeneity	AR(1)	α	β	$1 - \alpha - \beta$	1st Qu.	Median	3rd Qu.
Hom. Boot	No	Yes	0.00	0.00	1.00	0.025	0.030	0.035
Het. Boot	No	Yes	0.00	0.00	1.00	0.035	0.040	0.045
Perm	No	Yes	0.03	0.00	0.97	0.020	0.025	0.030
Hom. Boot	Yes	No	0.95	0.00	0.05	0.005	0.008	0.010
Het. Boot	Yes	No	0.00	0.01	0.98	0.260	0.330	0.480
Perm	Yes	No	0.04	0.00	0.95	0.245	0.325	0.452
Hom. Boot	Yes	Yes	0.94	0.00	0.06	0.005	0.013	0.015
Het. Boot	Yes	Yes	0.01	0.01	0.98	0.270	0.370	0.465
Perm	Yes	Yes	0.04	0.00	0.96	0.245	0.365	0.440

Table 7: Power for methods, updated (3/5). Note that with the slope being 0.25, nearly all methods detect difference within interval. A better estimate of type II error is given when slope = 0.025

The results in Table 7 are a summary of all of the methods found by taking the mean of each of the results presented. This is intended to interrogate the performance of each of these methods when underlying assumptions are unknown or unspecified. We find a robust performance for each of the new methods presented, maintaining a reasonable relationship between FWER and power. The metrics associated with homogeneous bootstrap are perhaps a bit misleading here as they appear to demonstrate exceptional power, though at the cost of unacceptable type I error.

Method	α	β	$1 - \alpha - \beta$	1st Qu.	Median	3rd Qu.
Hom. Bootstrap	0.630	0.000	0.370	0.012	0.017	0.020
Het. Bootstrap	0.002	0.008	0.990	0.188	0.247	0.330
Permtuation	0.035	0.000	0.965	0.170	0.238	0.307

Table 8: Summary of methods for Type II error

4.1.1 Summary of methods

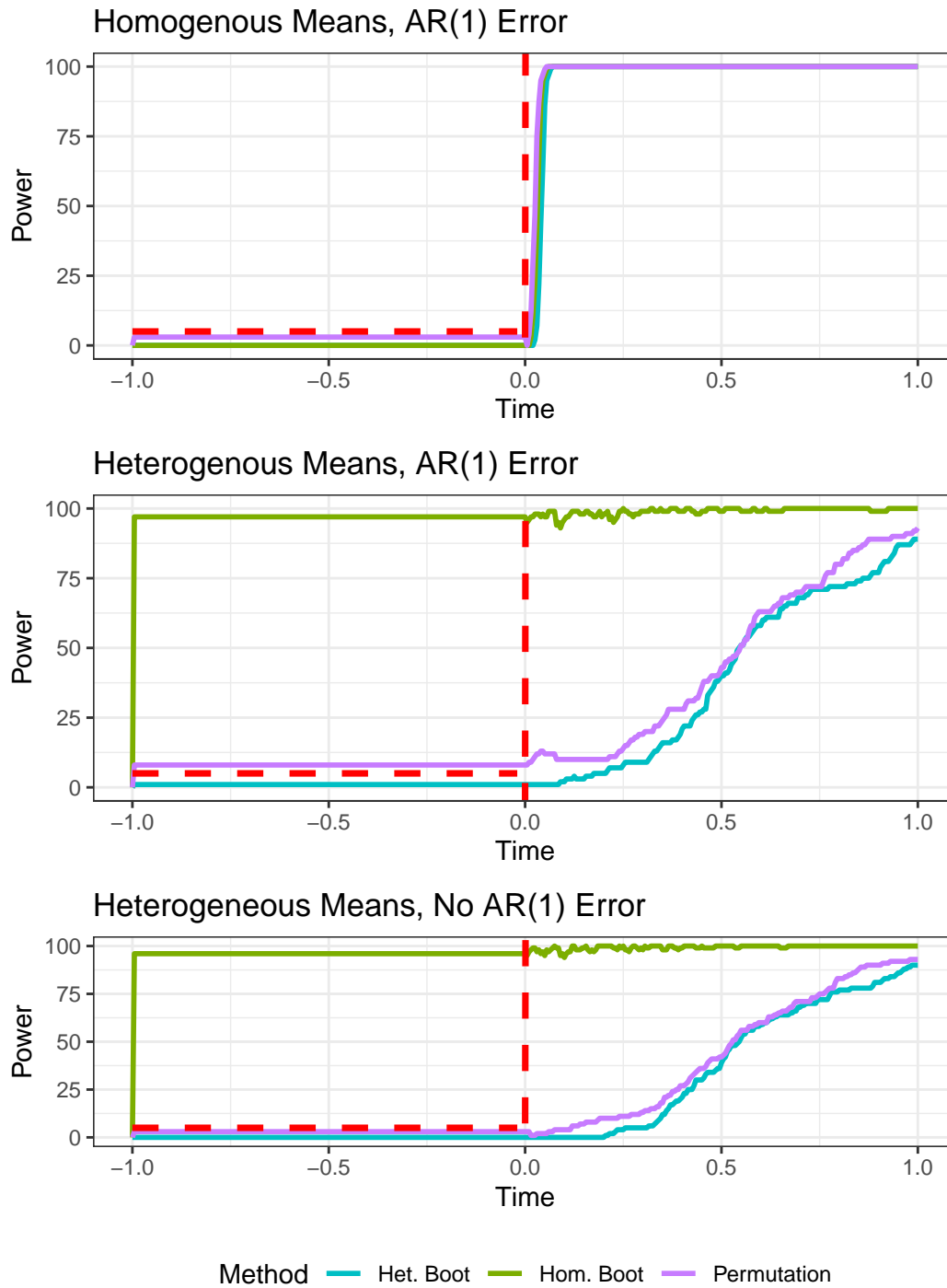


Figure 2: Observed power of each of the methods at each time in $(-1,1)$ (not sure if better as subplots?)

5 Discussion and concluding remarks

We set out both to interrogate the validity of the homogeneous bootstrap assumptions and to propose two alternative methods that would be more robust under a greater variety of assumptions. In doing so, we demonstrated conclusively the utility of the heterogeneous bootstrap and permutation tests while also highlighting a major shortcoming of the original. It's worth noting, however, that the FWER adjustment proposed in ? is still valid, if not slightly conservative, and with power similar to that of the permutation method.

In light of the results presented, one issue of concern is addressing the fact that a version of `bdots` with the homogeneous mean assumption was presented in 2018 and remained accessible on CRAN until the end of 2022. This has implications for the number of papers in which `bdots` may have demonstrated significance between groups when the underlying assumptions of homogeneous mean structure did not hold, as is likely the case in all instances related to the VWP. Concurrent with this issue is the issue of identifying current users of `bdots` of this change, as results found only a month ago will be profoundly different than what is seen today. At present, I am not sure the best way to address either of these. In either case, however, it will be prudent to remove this option from the `bdots` package all together, as there appears to be no obvious advantage to the homogeneous bootstrap over the others in terms of either controlling the FWER or obtaining power, even when the homogeneous mean structure assumption is met.

There are several limitations of the current paper that are worthy of further investigation. First, First, limited consideration was given to the effect of sample density on the observed type I error rate or power. As the fitting function in `bdots` simply returns a set of parameters, one could conceivably perform any of the methods presented on any arbitrary collection of points, whether or not any data were observed there. This extends itself to the condition in which subjects were sampled at heterogeneous time points, as may be the case in many clinical settings. What impact this may have or how to best handle these cases remains open for exploration. It is also worth investigating in greater detail what impact the re-drawing of subject specific parameters from their respective distributions has on both the FWER and power, as in several of the simulations the observed FWER was much lower than the nominal level. Particularly in the case of the permutation method which is *not* seeking to estimate the group distributions, it may be worthwhile to see if a favorable trade can be made to increase the resulting power.

We conclude by noting that `bdots` is now equipped with two methods to effectively control the FWER when assessing the differences in time series under a greater set of underlying assumptions, including those involving the presence of highly correlated test statistics. Further, both methods presented are robust to misspecification of the error structure while maintaining an acceptable FWER and adequate power.

6 Appendix – Full Power Simulations

Here we present the full collection of power sims, which, in addition to those given in Table 6 includes cases for heterogeneous means where autocorrelation is specified when fitting with `bdots`. This is indicated in the “AR(1) Specified” column. Additionally, plots giving the power at each time are presented in Figure 3.

Notably from this table, we see that in the case in which the true errors are IID, there is no measurable effect on power when an autocorrelated structure is incorrectly specified. This is similar to the opposite situation, in which the true error does have an AR(1) structure. In this case, we observe a marginal benefit to correctly specifying an AR(1) structure. This may in fact make retaining the AR(1) assumption a reasonable default in the `bdots` package.

We conclude by noting that of all of the methods investigated, that using permutation testing seems to have the preferable balance between controlling the FWER and power.

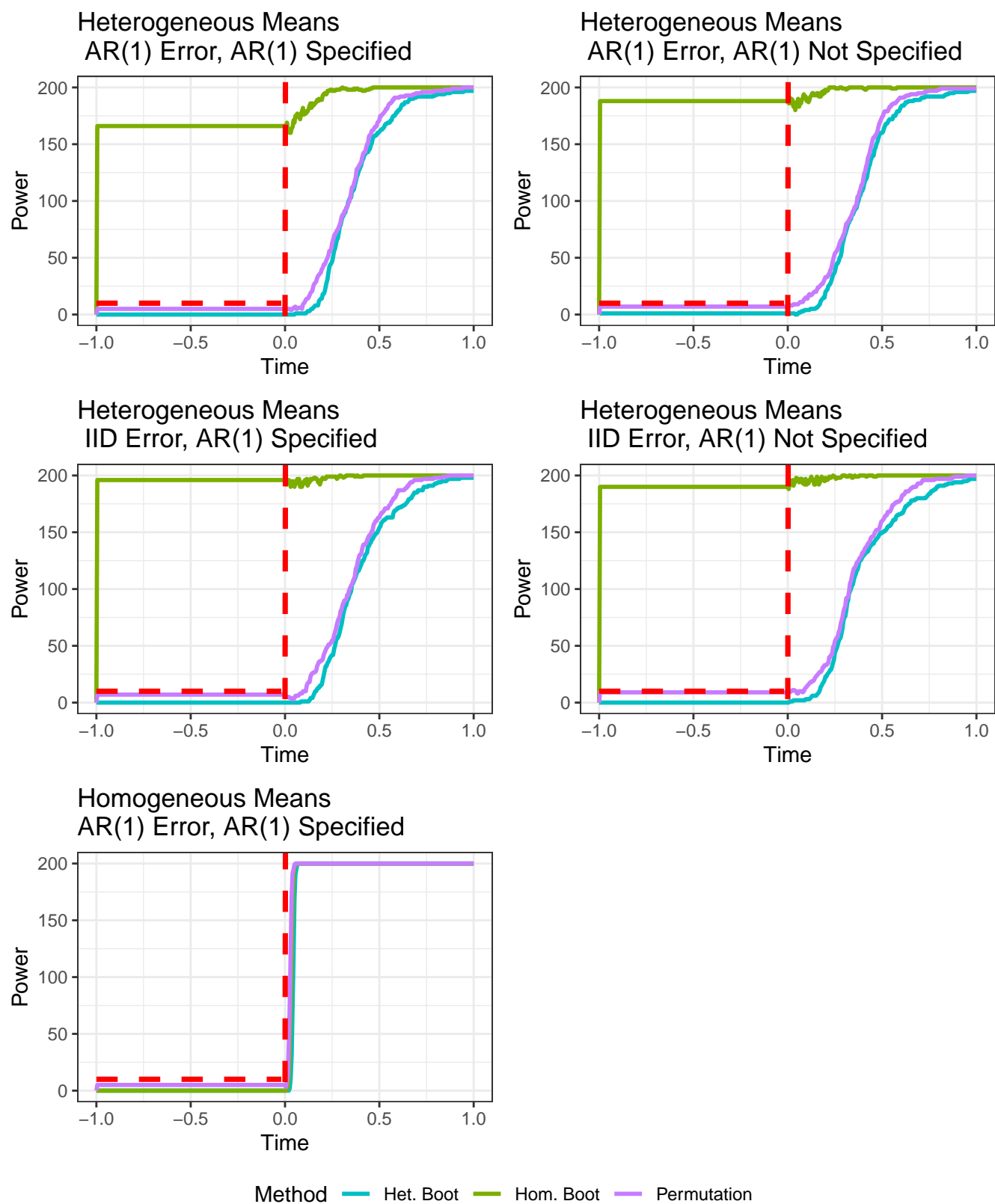


Figure 3: Power plots in time for each of the simulation settings. Note that in the heterogeneous means case, there is little difference when AR(1) is incorrectly specified

Method	Heterogeneity	AR(1) Error	AR(1) Specified	α	β	$1 - \alpha - \beta$	1st Qu.	Median	3rd Qu.
Hom. Boot	No	Yes	Yes	0.00	0.00	1.00	0.025	0.030	0.035
Het. Boot	No	Yes	Yes	0.00	0.00	1.00	0.035	0.040	0.045
Perm	No	Yes	Yes	0.03	0.00	0.97	0.020	0.025	0.030
Hom. Boot	Yes	No	No	0.95	0.00	0.05	0.005	0.008	0.010
Het. Boot	Yes	No	No	0.00	0.01	0.98	0.260	0.330	0.480
Perm	Yes	No	No	0.04	0.00	0.95	0.245	0.325	0.452
Hom. Boot	Yes	No	Yes	0.98	0.00	0.02	0.005	0.008	0.010
Het. Boot	Yes	No	Yes	0.00	0.01	0.99	0.261	0.350	0.475
Perm	Yes	No	Yes	0.04	0.00	0.96	0.225	0.335	0.440
Hom. Boot	Yes	Yes	No	0.94	0.00	0.06	0.005	0.013	0.015
Het. Boot	Yes	Yes	No	0.01	0.01	0.98	0.270	0.370	0.465
Perm	Yes	Yes	No	0.04	0.00	0.96	0.245	0.365	0.440
Hom. Boot	Yes	Yes	Yes	0.83	0.00	0.17	0.021	0.032	0.040
Het. Boot	Yes	Yes	Yes	0.00	0.01	0.98	0.250	0.330	0.450
Perm	Yes	Yes	Yes	0.03	0.00	0.97	0.223	0.335	0.428

Table 9: Power for methods