With great power comes greater responsibility:

cheating with bdots

Abstract

Gutting this entire thing and starting from scratch

1 Introduction

Free write

The original bdots presented a method whereby the difference in time series between two groups could be analyzed using a FWER correction to the alpha based on the autocorrelation of the resulting t-statistics

While we can confirm that the published results do indeed hold for the case they presented, such a situation is likely atypical and not reflective of the typical case involving VWP data. bdots involved testing between two groups, yet assumed that all subjects within a group had identical mean structures – that is, there was no between-subject variability to be accounted for.

More likely it is the case that in the comparison of two groups, each group has a typical distribution of parameters for its subjects. God its annoying a little bit how bad and out of order this is, but free write so its ok for now. As we show, when the initial (and restrictive) assumptions made in the original bdots doesn't hold, the resulting TIE is beyond what would be acceptable in most cases.

What we present instead is two alternatives, accommodating flexibility in two of the assumptions made in the original bdots. First, we propose a modified bootstrapping procedure that adequately accounts for observed between subject variability while retaining the novel FWER adjustment method presented for autocorrelated errors. In addition to this, we offer a play on a standard permutation test between the 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 by describing the two proposed alternatives to the original bdots bootstrap. We then outline the details of the simulations in demonstrating the TIE rate across a number of experimental conditions, along with the results. Finally if there is time we consider

a power analysis of the two resulting methods. I guess it's also fine to consider power in the case in which the original bdots does well. Naturally, the power is great, but then so what?

2 Detail on the original

Most generally the original bdots proposed that we have empirically observe data resulting from some mean structure with an associated error, with

$$y_{it} = f(\theta_{it}) + \epsilon_{it} \tag{1}$$

where

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

Under this paradigm, the errors could be iid normal (with $\phi = 0$) or have an AR(1) structure, with $0 < \phi < 1$. The unspoken assumption, however, with two subjects from the same group is that $\theta_{it} = \theta_{jt}$ for all i, j. In other words, it was assumed that there was no variability in the mean structure between subjects in the same group. This is also evidenced in the original bdots algorithm:

1. For each subject, fit the nonlinear function, specifying 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 approximate sampling distributions in (1.), randomly draw one bootstrap estimate for each of the model parameters on every subject
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 individuals
4. Use the mean estimates to determine the predicted population level curve, which provides the average population response at each time point

The previous statement is demonstrated in step (2.), where each subject is included in each iteration of the bootstrap.

Maybe here (to tie into the bottom) I could note that $\theta_i \sim N(\theta, V)$ except here $V = 0$

3 Proposed Methods

Here, we will describe in detail each of the proposed methods

3.1 Modified Bootstrap

A more likely case involving subjects in the VWP (or subjects within any group exhibiting between and within subject variability) is as such: suppose for example that we are considering a family of four parameter logistic curves, defined

$$f_\theta(t) = \frac{p-b}{1+\exp\left(\frac{4s}{\mathrm{p}-b}(x-t)\right)} + b \tag{3}$$

where $\theta = (p, b, s, x)$, the peak, baseline, slope, and crossover parameters, respectively. The distribution of parameters for subjects within this group may be normally distributed, with any individual subject i's parameters following the distribution

$$\theta_i \sim N(\mu_\theta, V_\theta). \tag{4}$$

In the course of collecting observed data on subject i, we may find that there is a degree of variability in our observations between trials, reflected in the standard errors derived when fitting the observed data to the functional form in Equation 3. This gives us a distribution for the observed parameter,

$$\hat{\theta}_i \sim N(\theta_i, s_i^2). \tag{5}$$

This is where I perhaps don't have my notation quite how I want it (what do i know about bayesian things or hierarchical models? besides, im a man, i derive my statistics from the gut). We also need to be clear on language. We have a few things here:

1. It's not really a bootstrap when we sample $\theta^*_{ib} \sim N(\hat{\theta}_i, s_i^2)$. It's the bth estimate of θ_i following distributoin just given
2. If we sample *without* replacement and take the mean, we essentially have a sum of independent normals (start notation from efron/tibshirani maybe will use it maybe not)

$$\theta^*_b = \frac{1}{n}\sum \theta^*_{ib}, \quad \theta^*_b \sim N\left(\mu_\theta, \frac{1}{n^2}\sum s_i^2\right) \tag{6}$$

   but note that this isn't *really* a bootstrap of the θ_i values. I'm not really sure what you would call this. Maybe we shouldn't have the b subscript but call it something else
3. Alternatively if we sample *with* replacement, we still have an individual draw for each bootstrapped subject, $\theta^*_{ib} \sim N(\hat{\theta}_i, s_i^2)$, but now the distribution of the for-real bootstrapped parameter is

$$\theta^*_b \sim N\left(\mu_\theta, \frac{1}{n}V_\theta + \frac{1}{n^2}\sum s_i^2\right) \tag{7}$$

   The mean value across all bootstraps will be the same as before, but now we are actually accounting for the variability that exits.

4. As an aside, this sets us up for a useful callback – when describing the mean structure of the simulations, we can say that they both follow $\theta \sim N(\mu_\theta, V_\theta)$, but one has empirically determined V_θ and the other sets $V_\theta = 0$. We can then be like hey go look at Equations 6 and 7 and see how the correct bootstrap can accomodate both cases but bad bootstrap cannot

5. Last aside – I have not simulated this yet, but Bob did make a comment about the TIE for the good bootstrap being too low, asking about sacrifice in power. While we haven't done power yet (as of this writing), I wonder how the good bootstrap would look if we disregarded the s_i^2 terms

—

We also note (and verify in the simulations) that it is not always necessary to specify an AR(1) autocorrelation structure to the errors in the model. While failing to include it slightly inflates the TIE error when the data truly is autocorrelated, when the data is not it can lead to overly conservative estimates. As such, we make the propose the following changes to the original bootstrap algorithm:

1. In step (1.), the specification of AR(1) structure is *optional* and can be modified with arguments to functions in `bdots`

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.

A paired bootstrapping can be implemented by performing this same algorithm but ensuring that at each iteration of the bootstrap the same subjects are sampled with replacement in each group. This happened by default in the original implementation as each subject was retained in each iteration of the bootstrap.

3.2 Permutation 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:

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 $\overline{p}_{jt}$ and s^2_{jt} for $j = 1, 2$. Use these values to compute a test statistic T_t at each time point,

$$T_t = \frac{|\overline{p}_{1t} - \overline{p}_{2t}|}{\sqrt{s^2_{1t} + s^2_{2t}}}. \tag{8}$$

   This will be our observed test statistic.

3. Repeat (2) P additional times, each time shuffling the group membership between subjects. This time, we fitting 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, each time retaining the maximum value. This collection of P statistics will serve as our null distribution which we denote $\widetilde{T}$ (or whatever we wanna call it). Let $\widetilde{T}_\alpha$ be the 1 - $\alpha/2$ quantile of $\widetilde{T}$
4. Compare each of the observed T_t with $\widetilde{T}_\alpha$. Areas where $T_t > \widetilde{T}_\alpha$ are designated significant.

Paired permutation testing is implemented with a minor adjustment to step (3). Instead of permuting all of the labels between groups, choose one group and randomly assign each subject to either retain their current group membership or to change groups. Make the corresponding reassignment to members in the second group. .This ensures that each permuted group contains one observation from each subject.

4 TIE Simulations

this section needs reorganized but it is as is for now

When now go about comparing the type I error rate of the three methods just described. In doing so, we will establish several conditions under which the observed subject data may have been generated or fit. This includes two conditions for the mean structure, two conditions for the error structure, paired and unpaired data, and data fit with and without an AR(1) assumption. Considering each permutation of this arrangement results in sixteen different simulations. Each simulation will then be examined for type I error using each of the three methods described (I said that twice).

Each set of conditions generates two groups, with $n = 25$ subjects in each group, with $N = 100$ simulated trials for each subject. Each simulation was conducted 100 (1,000 running) times to determine the rate of type I error. And as this is an examination of the type I error rate, both of the two groups compared were constructed using the same distributions and manners described

4.1 Data Generation

Data was generated according to three conditions: mean structure, error structure, and paired status. We assume that each subject's data was of the general form

$$y_{it} = f_{\theta_i}(t) + \epsilon_{it}. \tag{9}$$

We assume that each group drew subject-specific parameters from a normal distribution,

$$\theta_i \sim N(\mu_\theta, V_\theta). \tag{10}$$

In all simulations, we set $\mu_\theta = (0, 0.8, 0.002, 750)$ (it actually was not this, but I need to go look at what it was), corresponding to the baseline, peak, slope, and crossover parameters from Equation 3. In half of our simulations, we set V_θ was determined following an empirical distribution from (Timbell 2017), giving an estimate of the typical amount of variability observed among normal hearing subjects. In the remaining cases, we set $V_\theta = 0$, a silly idea, sure, but assuming that each of the subjects' observations is derived from the same mean structure, with differences only in the observed error.

The error structure was of the form

$$e_{it} = \phi e_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma) \tag{11}$$

where the w_{it} are iid with $\sigma = 0.025$ (it actually is something else that is variable depending on number of trials, but such that for 100 trials it is equal to this, same as Oleson 2017). 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 iid.

Finally, we consider the paired data, which differs in creation according to the mean structure. In the case in which $V_\theta \neq 0$, we simply used the same value of θ_i for the ith subject in each group, allowing the only difference to be that corresponding to the error structure. In the case when $V_\theta = 0$, however, it was already the case that the set of parameters were the same between subjects in each group (and indeed for all subjects in both groups). As such, letting the observed data for subject i in group A be denoted y_{iA}, we set

$$y_{iB} = y_{iA} + N(0, \sigma)$$

so that the only difference between paired subjects was uncorrelated normal noise at each time point. (I get that this is questionable, but what is the alternative? Truly it would just be to half the degrees of freedom in the t-test, which isn't really comparing IRL paired data)

4.2 extra/should i elaborate on this instead of the general summary above?

Not sure if this is worth discussing here or in some section detailing the actual construction of these estimates – what is actually done is we for a subject with N trials at each time point we have

$$y_{it} \sim Bin(N, f_{\theta_t}(t)) \tag{12}$$

This has the nice benefit of creating normally distributed errors with variances modulated by the number

of trials. In the other case where it is added separately, we set $\sigma = (1-p)(p)/\sqrt{N}$ so that if we do change the number of observed trials, we are able to adjust the variability in both cases. Pretty slick, i know

Actually, I had previously gone into this level of detail for both mean structure and error structure. Doing so kind of necessitated breaking these into longer, independent sections. As it is now, I kind of walk through all of it in a paragraph or two (above). I'm not really sure how much the detail/length trade off matters here as i imagine most people wont care about the particulars

5 Type I Error

The parameters used in the group distributions were empirically determined from (timball 2017) and set $b = 0.21$, $p = 0.90$, $s = 0.00165$ and $x = 725.03$ (how many decimals, idk. Also won't include variability but I have that too). A four parameter logistic curve with these parameters has the following form:

5.1 FWER

Here are the results for type I FWER (I have put the new values for permutation in parenthetical, this is after accounting for the within-subject variance with the additional drawing step for coefficients. I only tested this on the first four in unpaired, the rest are not updated)

manymeans	ar1	bdotscor	Bad Bootstrap	Good Bootstrap	Permutation (updated)
FALSE	TRUE	TRUE	0.06	0.01	0.21 (0.05)
FALSE	TRUE	FALSE	0.87	0.08	0.18 (0.11)
FALSE	FALSE	TRUE	0.08	0.00	0.14 (0.03)
FALSE	FALSE	FALSE	0.15	0.02	0.21 (0.04)
TRUE	TRUE	TRUE	0.92	0.03	0.03
TRUE	TRUE	FALSE	0.96	0.02	0.04
TRUE	FALSE	TRUE	0.99	0.05	0.01
TRUE	FALSE	FALSE	1.00	0.05	0.03

Table 1: TIE for realistic parameters (unpaired)

manymeans	ar1	bdotscor	Bad Bootstrap	Good Bootstrap	Permutation
FALSE	TRUE	TRUE	0.12	0.02	0.03
FALSE	TRUE	FALSE	0.86	0.08	0.03
FALSE	FALSE	TRUE	0.09	0.01	0.01
FALSE	FALSE	FALSE	0.14	0.01	0.03
TRUE	TRUE	TRUE	0.49	0.02	0.01
TRUE	TRUE	FALSE	0.94	0.03	0.02
TRUE	FALSE	TRUE	0.72	0.02	0.00
TRUE	FALSE	FALSE	0.74	0.04	0.00

Table 2: TIE for realistic parameters (paired)

5.2 Median per comparison error rate

manymeans	ar1	bdotscor	Bad Bootstrap	Good Bootstrap	Permutation (updated)
FALSE	TRUE	TRUE	0.01	0.00	0.04 (0.00)
FALSE	TRUE	FALSE	0.31	0.00	0.04 (0.02)
FALSE	FALSE	TRUE	0.00	0.00	0.02 (0.00)
FALSE	FALSE	FALSE	0.00	0.00	0.03 (0.00)
TRUE	TRUE	TRUE	0.51	0.01	0.01
TRUE	TRUE	FALSE	0.76	0.01	0.00
TRUE	FALSE	TRUE	0.86	0.01	0.00
TRUE	FALSE	FALSE	0.81	0.01	0.00

Table 3: median per comparison error rate (unpaired)

manymeans	ar1	bdotscor	Bad Bootstrap	Good Bootstrap	Permutation
FALSE	TRUE	TRUE	0.03	0.00	0.00
FALSE	TRUE	FALSE	0.26	0.00	0.00
FALSE	FALSE	TRUE	0.00	0.00	0.00
FALSE	FALSE	FALSE	0.01	0.00	0.00
TRUE	TRUE	TRUE	0.13	0.00	0.00
TRUE	TRUE	FALSE	0.52	0.02	0.00
TRUE	FALSE	TRUE	0.38	0.01	0.00
TRUE	FALSE	FALSE	0.44	0.01	0.00

Table 4: median per comparison error rate (paired)

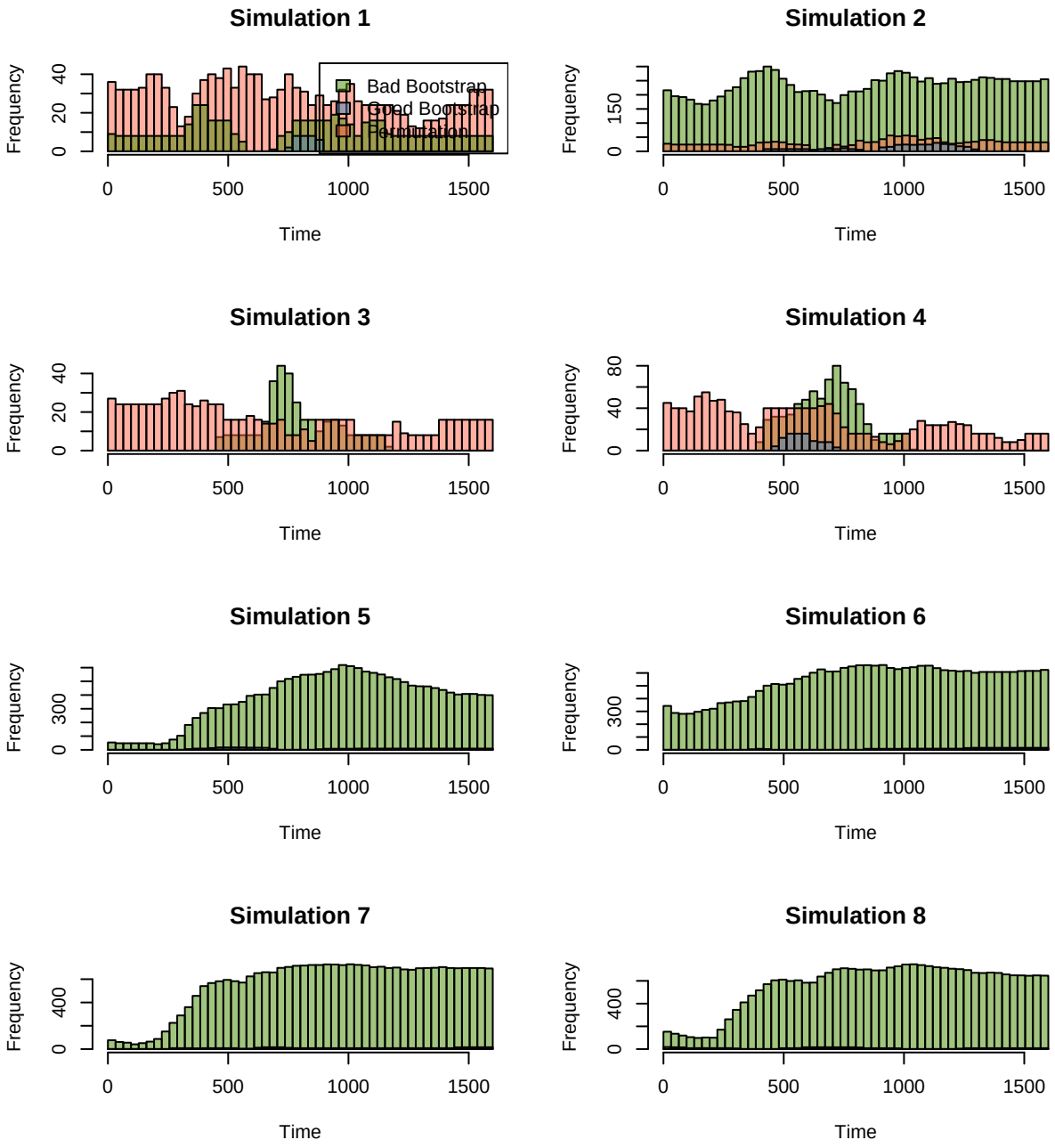


Figure 1: quick put together of what this might look like, frequency of TIE at each spot. obviously the plots cant stay like this. potentially misleading looking at absolute frequency, but this is consequence of breaks being length 32 (or else you cant see color). Just did not take time to do in ggplot. Also, haven't labeled tables yet but this corresponds to the first 8 sims in unpaired. I would maybe like this plot more if the y axis was percentage

6 Power Simulations

A few notes to qualify what follows:

1. Permutation has been updated for power, so it correctly samples from the gnls distribution for each subject
2. We should discuss what simulations to include bad bootstrap in. I have included them all here. And to prevent it from just flagging entire regions as different, I basically generated the data to be "paired" with regards to the intercept parameter, meaning for half of the generated data, its pretending likes its not paired when it is. There are still a few cases where TIE occurs, and I have noted them and removed them from the summary
3. There are a few different simulation results presented here, all different in slightly different ways (time intervals, variability in generating distribution, etc). I have tried to note which is which
4. they might want paired testing

All this talk on the type I error rate sure is interesting, but what good is having a low type I rate if we are just trading it in for type II? In this hard hitting piece of investigative journalism, we set out to determine the empirical power of the proposed methods under a variety of conditions, similar to those above but excluding the case of paired observations. Inb4 comments like "oh but power between paired cases is what we are interested in most!", well, maybe if there's time.

To determine power, groups were simulated from two mean structures of the form

$$y = \begin{cases} b & x < 0 \\ mx + b & x \geq 0 \end{cases} \tag{13}$$

where $b \sim |N(0, 0.005)|$ and $m \sim N(\mu_i, 0.005)$, where $\mu_i = 0$ for the group without effect and $\mu_i = 0.5$ for the group with effect (I guess this is called a folded normal distribution? I could have written something like $b = |\gamma|, \gamma \sim N(0, 0.005)$, but maybe there is an even slicker way of expression this). A depiction of these mean structures is given in Figure 2.

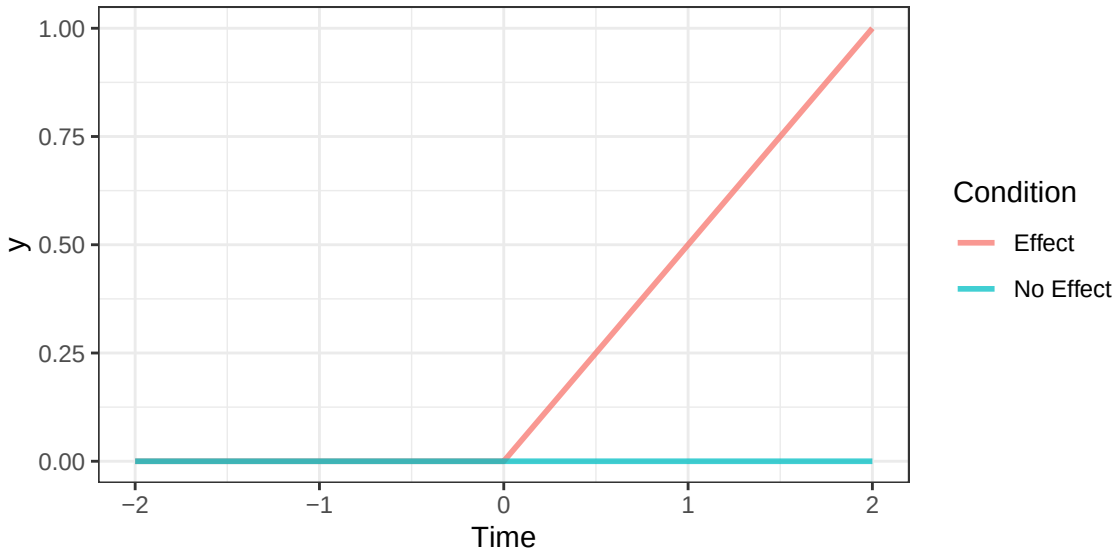


Figure 2: Plot of the generating mean structure for power simulations. This looks more drastic than it is because the x axis spans 4 while the y only 1

6.1 Data generation – WARNING: WIP

As before, we tested a limited combination of "manymeans", "ar1", and "bdotscor" (i know these need different names). In each simulation of (100) simulations, two groups were simulated, with 25 subjects in each group. We begin with a description of data generation for the manymeans=TRUE assumption. [also, how to avoid ambiguity with a statement like "this simulation (with these settings) had 100 simulations (instances of those settings)?]

Beginning with the "Effect" group, we draw slope and intercept parameters for each subject, taking the absolute value of each to ensure the slope is oriented in the correct direction. Fitting a mean structure to each subject's parameters, we then add error according to the AR(1) assumption identically to the method used in the calculation of the type I error rate.

For the "No Effect" group, we took as intercept parameters the same values drawn for the "Effect" group – this was to minimize the difference in distribution of the values in the range (-2, 0) since the TIE rate there is awful and not doing this resulted in significant differences everywhere (since it is constant in that region, having one difference means having them all different).

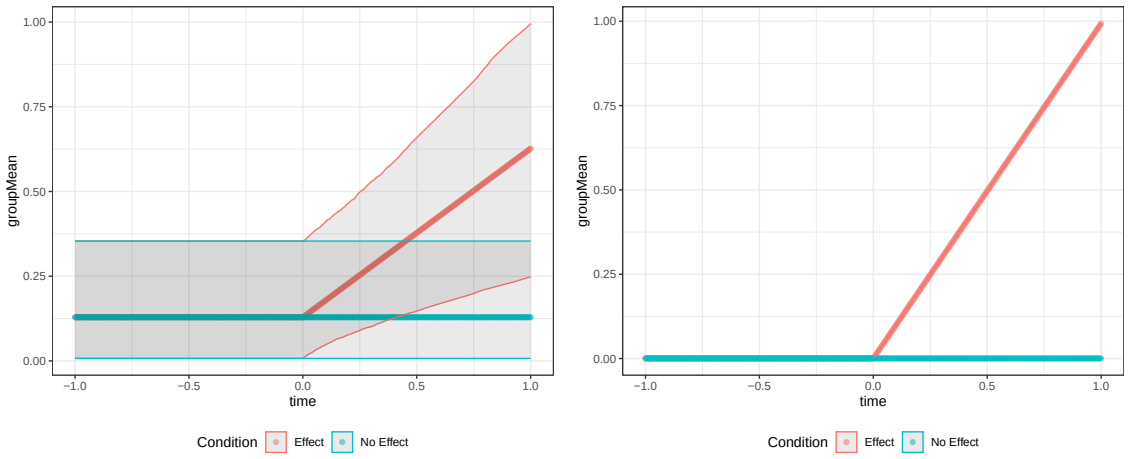
For the manymanys = FALSE, we basically did the same thing as above, but only drew parameters for one subject, assigning the remaining 24 subjects the same mean structure but deriving for each their own error structure. And actually, its a slight variant of that to help the sim along – I still drew 25 but then

picked the min/max of b and m to be closest to 0 and 0 or 0.5 depending on their groups, respectively. Is that cheating? I could also do like I did with the TIE sims and just set them to all be exactly 0 and 0/0.5 and have the error be the only difference

As a final thing, I also ran this over two different intervals: `seq(-1,1,length.out = 401)` and `seq(-2,2, length.out = 501)` to see if anything would change. Fortunately, the impact was not very noticeable. Alright, let's get to results

6.2 Results

New: I generated 1000 subjects and have plotted here the distribution of the *true* parameters, i.e., the generating distributions



(a) (b)

Figure 3: Paired bases, $\sigma = 0.025$ (a) Many means assumption, (b) single means distribution (obviously it's zero variability since one mean). Note also that the mean slope appears closer to one. This isn't surprising since the *actual* distribution takes the max slope of 25. I'll change this in future sims

The columns for manymeans, ar1, and bdotscorr are the same as before. Right now, Time = 2 indicates from (-2,2) while Time = 1 is (-1,1). This table presents a summary of the time at which a difference was detected. In all cases, the true effect begins at 0. It has $\sigma = 0.025$, time for (-1,1) and (-2,2), the intercept term is "paired", everything else should be certified kosher. 500 bootstraps/permutations, 100 sims for each

manymeans	ar1	bdotscorr	Time	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	1	0.0000	0.0050	0.0100	0.0100	0.0107	0.0100	0.0200
FALSE	TRUE	TRUE	2	0.0000	0.0080	0.0080	0.0160	0.0121	0.0160	0.0160
TRUE	TRUE	FALSE	1	0.2500	0.0050	0.0050	0.0050	0.0070	0.0100	0.0100
TRUE	TRUE	FALSE	2	0.0000	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080
TRUE	FALSE	FALSE	1	0.0000	0.0050	0.0050	0.0050	0.0051	0.0050	0.0100
TRUE	FALSE	FALSE	2	0.0000	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080

Table 5: Power for bad bootstrap

manymeans	ar1	bdotscorr	Time	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	1	0.0000	0.0100	0.0100	0.0150	0.0143	0.0150	0.0250
FALSE	TRUE	TRUE	2	0.0000	0.0080	0.0160	0.0160	0.0155	0.0160	0.0240
TRUE	TRUE	FALSE	1	0.0000	0.0950	0.1300	0.1450	0.1457	0.1650	0.2050
TRUE	TRUE	FALSE	2	0.0000	0.1440	0.1460	0.1480	0.1480	0.1500	0.1520
TRUE	FALSE	FALSE	1	0.0000	0.0850	0.1300	0.1400	0.1433	0.1562	0.2150
TRUE	FALSE	FALSE	2	0.0000	0.1040	0.1360	0.1440	0.1476	0.1600	0.2000

Table 6: Power for good bootstrap

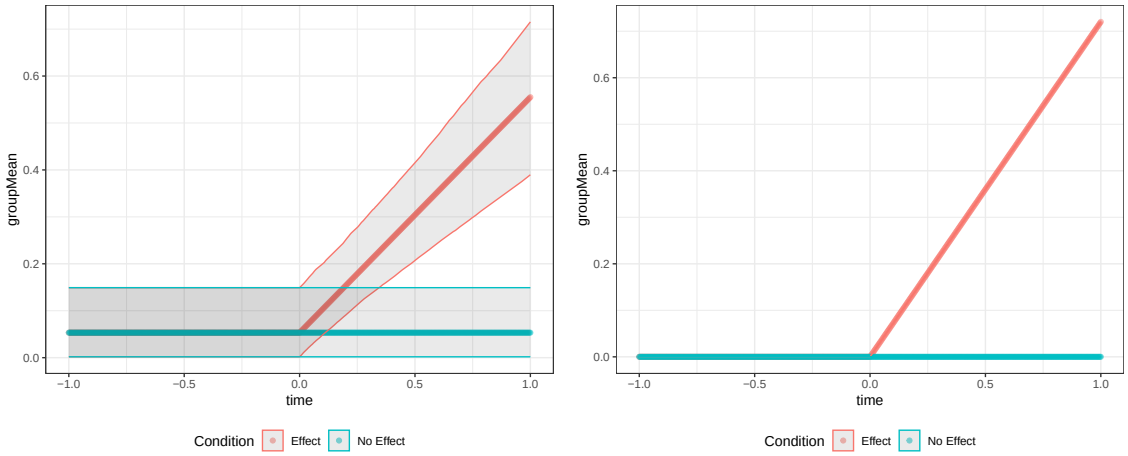
manymeans	ar1	bdotscorr	Time	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	1	0.0100	0.0050	0.0100	0.0100	0.0094	0.0100	0.0200
FALSE	TRUE	TRUE	2	0.0000	0.0080	0.0080	0.0080	0.0101	0.0160	0.0160
TRUE	TRUE	FALSE	1	0.0000	0.0950	0.1300	0.1525	0.1529	0.1712	0.2150
TRUE	TRUE	FALSE	2	0.0000	0.1440	0.1440	0.1440	0.1440	0.1440	0.1440
TRUE	FALSE	FALSE	1	0.0000	0.0800	0.1300	0.1450	0.1472	0.1600	0.2250
TRUE	FALSE	FALSE	2	0.0000	0.0960	0.1280	0.1480	0.1475	0.1600	0.2080

Table 7: Power for permutation

6.3 smaller variance version

looking at times (-2, 2), leaving the baseline parameter "paired", but significantly reducing the variance of the distributions im drawing from ($\sigma = 0.005$). Also only did 50 simulations for each

Here is plot of distributions first (with lower variance)



(a) (b)

Figure 4: Paired bases, $\sigma = 0.005$ (a) Many means assumption, (b) single means distribution

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.0000	0.0080	0.0080	0.0160	0.0136	0.0160	0.0160
TRUE	TRUE	FALSE	0.2600	0.0080	0.0080	0.0080	0.0082	0.0080	0.0160
TRUE	FALSE	FALSE	0.0000	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080

Table 8: Power for bad bootstrap

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.0000	0.0080	0.0160	0.0160	0.0178	0.0160	0.0240
TRUE	TRUE	FALSE	0.0000	0.0480	0.0640	0.0800	0.0750	0.0800	0.1040
TRUE	FALSE	FALSE	0.0000	0.0560	0.0660	0.0720	0.0755	0.0800	0.1120

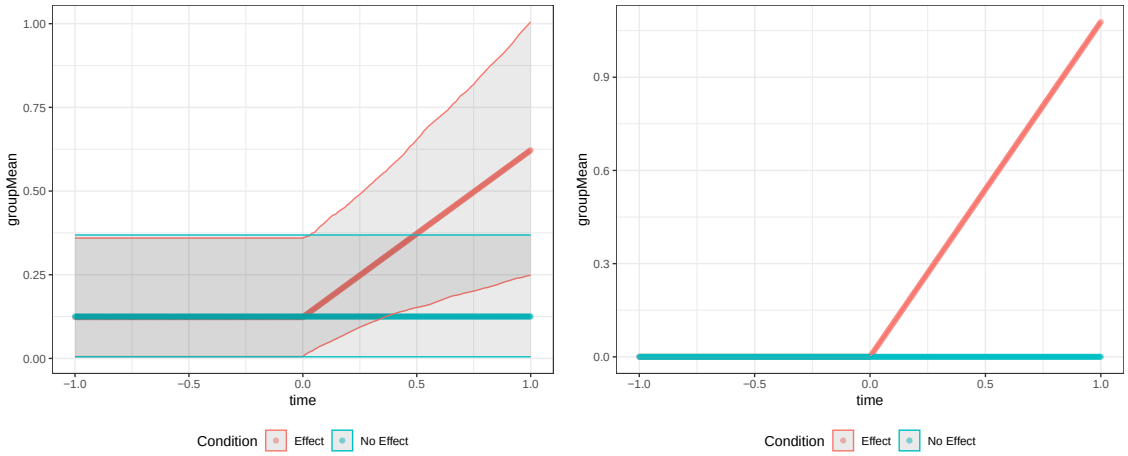
Table 9: Power for good bootstrap

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.0200	0.0080	0.0080	0.0160	0.0121	0.0160	0.0160
TRUE	TRUE	FALSE	0.0000	0.0400	0.0560	0.0720	0.0669	0.0720	0.0960
TRUE	FALSE	FALSE	0.0000	0.0480	0.0640	0.0640	0.0672	0.0720	0.0960

Table 10: Power for permutation

6.4 Trying with unpaired case

looking at times (-2, 2), making the baseline parameter "unpaired" and leaving variability at $\sigma = 0.025$. This is running now (8:15am) and should be finished by the time we meet. This has 100 for each sim



(a) (b)

Figure 5: Unpaired bases, $\sigma = 0.025$ (a) Many means assumption, (b) single means distribution

Some interesting things here, I think

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.4200	0.0080	0.0080	0.0120	0.0123	0.0160	0.0240
TRUE	TRUE	FALSE	0.9600	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080
TRUE	FALSE	FALSE	0.9200	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080

Table 11: Power for bad bootstrap

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.2600	0.0080	0.0080	0.0160	0.0157	0.0220	0.0320
TRUE	TRUE	FALSE	0.0000	0.0080	0.1040	0.1360	0.1496	0.1840	0.3440
TRUE	FALSE	FALSE	0.0000	0.0160	0.1040	0.1520	0.1589	0.1940	0.3760

Table 12: Power for good bootstrap

manymeans	ar1	bdotscorr	TIE rate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	TRUE	TRUE	0.4600	0.0080	0.0080	0.0080	0.0116	0.0160	0.0160
TRUE	TRUE	FALSE	0.0200	0.0160	0.0960	0.1360	0.1405	0.1760	0.2880
TRUE	FALSE	FALSE	0.0200	0.0240	0.0980	0.1520	0.1513	0.1920	0.3280

Table 13: Power for permutation

7 Discussion and concluding remarks

There isn't really a lot to say in this paper: we pretty quickly identified the issue with the assumption that $V = 0$ for the assumptions in the data generation parameters. We showed pretty conclusively that the type I error is absolutely bonkers in even the mildest of deviations from this, removing any justification from

this being the "default" method in `bdots`. With regards to fitting models assuming AR(1) error structure, there really wasn't any observed penalty in not doing so with the good bootstrap or permutation methods, and if anything it brought the observed TIE closer to 0.05, with some implications for power (though those simulations are still work in progress).

And speaking of power, here is the only case, really, where bad bootstrap can make any argument – it (naturally) detects a difference far beyond any of the other cases, though as we should discuss, this (so far) is because I have sort have cheated in how I created the data. Judging by the type I error rate, there may be only one case in which it's worth doing a power analysis at all. As to a comparison of the other two, preliminary results on the simulations I think are correct seem to ever so slightly favor permutations, which is neat, I kind of favor them. This is pending a further investigation in what sort of simulation here is really appropriate as far as data generation is concerned.

And that's pretty much it. Not a lot to discuss. There is an open question on the impact of time points (both quantity and range) on where and what we consider significant. Inflating the observed data with huge swaths of obviously different functions might be cheating. We may look at a few things in passing, but likely not enough to make any hard and fast conclusions. It's an interesting question, though, and especially when we consider comparing data with non-homogenous ranges, which we also didn't look it (i.e., mousey tumor data).

Appendix

Could include oleson 2017 parameters just to say we did it and verifying the two results that they had previously found (i.e., we have implemented this correctly). Commented out for now