

## Online Chapter 13 The Mass Univariate Approach and Permutation Statistics

### Overview

This chapter describes the mass univariate approach to the analysis of ERP data and the related permutation approach. The mass univariate approach is a powerful way of dealing with the problem of multiple comparisons. It was pioneered in neuroimaging research, where the problem of multiple comparisons is so obvious that it is always explicit rather than implicit. In early neuroimaging studies, investigators conducted a *t* test or ANOVA on each voxel separately, asking whether the activity in each voxel differed significantly across conditions. They then performed a Bonferroni correction for multiple comparisons. For an effect to be significant after a Bonferroni correction, the original *p* value must be less than the alpha value divided by the number of comparisons. For example, if you are testing the effect of some manipulation in 1000 individual voxels with a typical alpha of .05, a given voxel would need to have an uncorrected *p* value of  $.05/1000$  (.00005) to be considered significant. To get a significant effect after this draconian correction, the effect must be enormous! This makes it hard to find significant effects even when there are true differences between conditions (i.e., the Type II error rate is very high). Neuroimaging researchers sometimes call this “being Bonferronied to death.” The Bonferroni correction assumes that each of the tests is independent, which is not true of imaging data (or EEG/ERP data) because the values at adjacent variables (i.e., voxels) are highly correlated. Moreover, it imposes a very strict criterion for significance that is not always sensible. The result of these factors is very low statistical power.

Three different approaches have been developed that allow neuroimaging researchers to address the problem of multiple comparisons with improved statistical power (i.e., to achieve a true Type I error rate of .05 with a lower Type II error rate). All three can also be applied to the problem of multiple comparisons in ERP research, in which the problem arises from the number of time points and electrode sites rather than from the number of voxels. The first is to use the *false discovery rate* (FDR) correction instead of the Bonferroni correction. The second is to use *permutation statistics*, which use the actual structure of the recorded data to estimate the likelihood that an effect is reliable. The third is the *cluster-based* approach, in which the analysis looks for effects that form clusters over space and time (i.e., many consecutive time points and adjacent electrode sites that show the same pattern).

The mass univariate approach to ERP analysis was first popularized by Eric Maris and his colleagues (Maris, 2004; Maris & Oostenveld, 2007; Maris, 2012), and they also provided an implementation of it in a freely available Matlab toolbox, FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011). This implementation has been widely used, but it is not very user friendly. These methods are also now implemented in EEGLAB.

David Groppe has written a very accessible review of the mass univariate approach (Groppe, Urbach, & Kutas, 2011a), which is on my list of *Papers Every New ERP Researcher Should Read* (see the end of Chapter 1). David has also created a user-friendly implementation called the Mass Univariate Toolbox ([http://openwetware.org/wiki/Mass\\_Univariate\\_ERP\\_Toolbox](http://openwetware.org/wiki/Mass_Univariate_ERP_Toolbox)), which works in concert with EEGLAB and ERPLAB. The Mass Univariate Toolbox implements all three of the new approaches for controlling the Type I error rate while optimizing statistical power. Each has its strengths and weaknesses (which are nicely described by Groppe et al., 2011a). I will provide a relatively brief, nontechnical overview here.

### The Bonferroni Correction

I will begin by explaining the traditional Bonferroni correction. If you use the Bonferroni correction, the corrected *p* value is related to the likelihood that *any* of your effects is bogus. For example, imagine that we did a *t* test comparing the rare and frequent waveforms at each time point from 0-600 ms at every channel in the oddball experiment shown Figure 13.1 (which is the same as Figure 10.5 in Chapter 10). With a sampling rate

of 500 Hz, we would have 300 time points between 0 and 600 ms, and we would do a t test comparing rare and frequent at each of these time points at each of the 10 electrode sites, giving us a total of 3000 t tests. Without any correction, we would expect that approximately 150 (5%) of these t values would be significant by chance (i.e., if the null hypothesis were correct). This is why we need to do a correction for multiple comparisons. If we did a Bonferroni correction, a given comparison would be considered significant only if the uncorrected p value is less than .000017 ( $.05 \div 3000$ ). If the null hypothesis were true, there would be only a 5% chance that one or more of the 2500 t tests would yield a significant corrected p value (an uncorrected p of  $< .000017$ ). This is called *strong control* of the Type I error rate, because we can have confidence in every effect that is classified as significant.

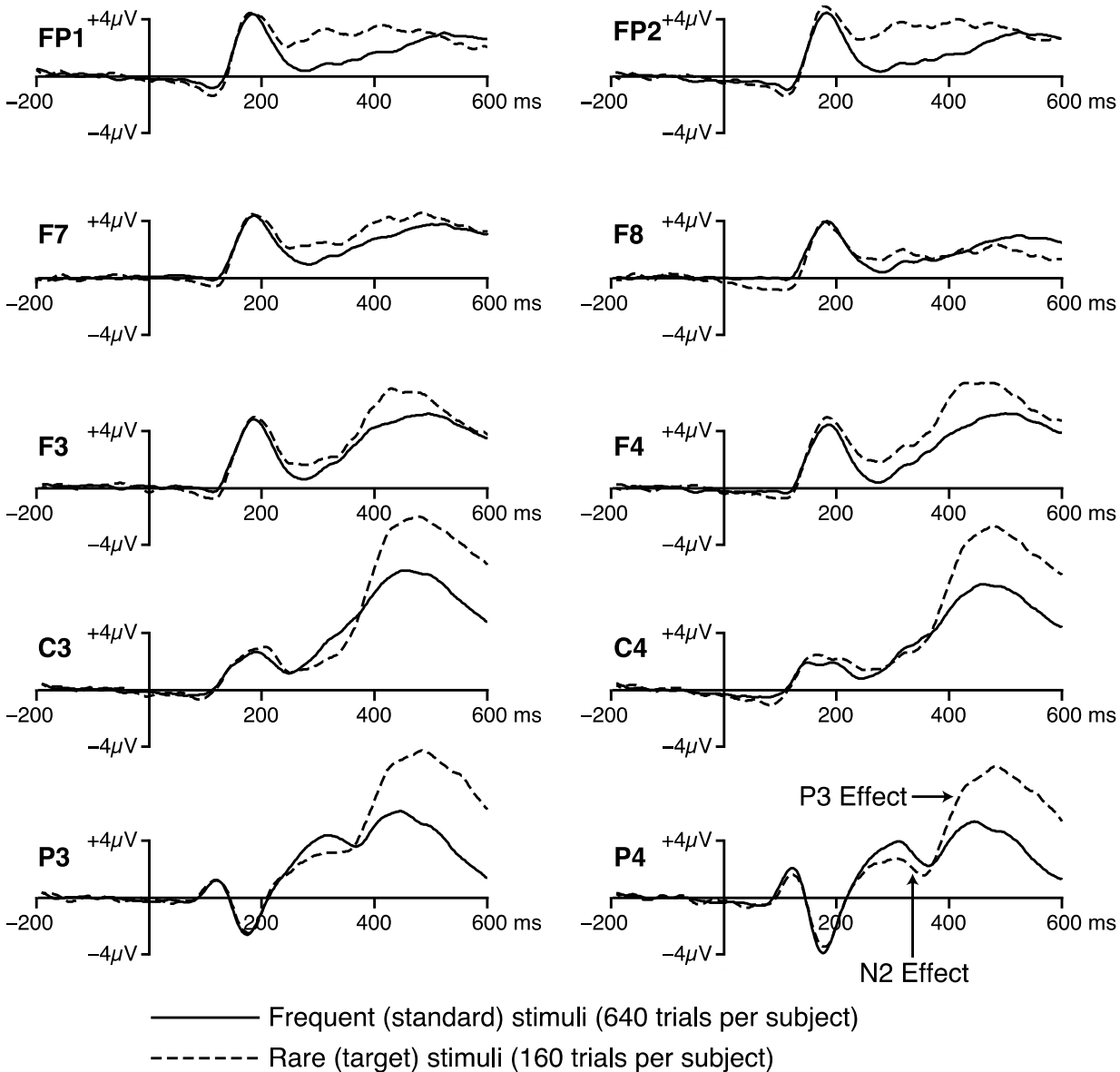


Figure 13.1. Grand average ERP waveforms from an example oddball experiment in which 20% of the stimuli were letters and 80% were digits (or vice versa, counterbalanced across trial blocks). The data are from a subset of 12 of the control subjects who participated in a published study comparing schizophrenia patients with control subjects (Luck et al., 2009). These data are referenced to the average of the left and right earlobes and were low-pass filtered offline (half-amplitude cutoff = 30 Hz, slope = 12 dB/octave). Note that this is the same as Figure 10.5 in the main body of Chapter 10.

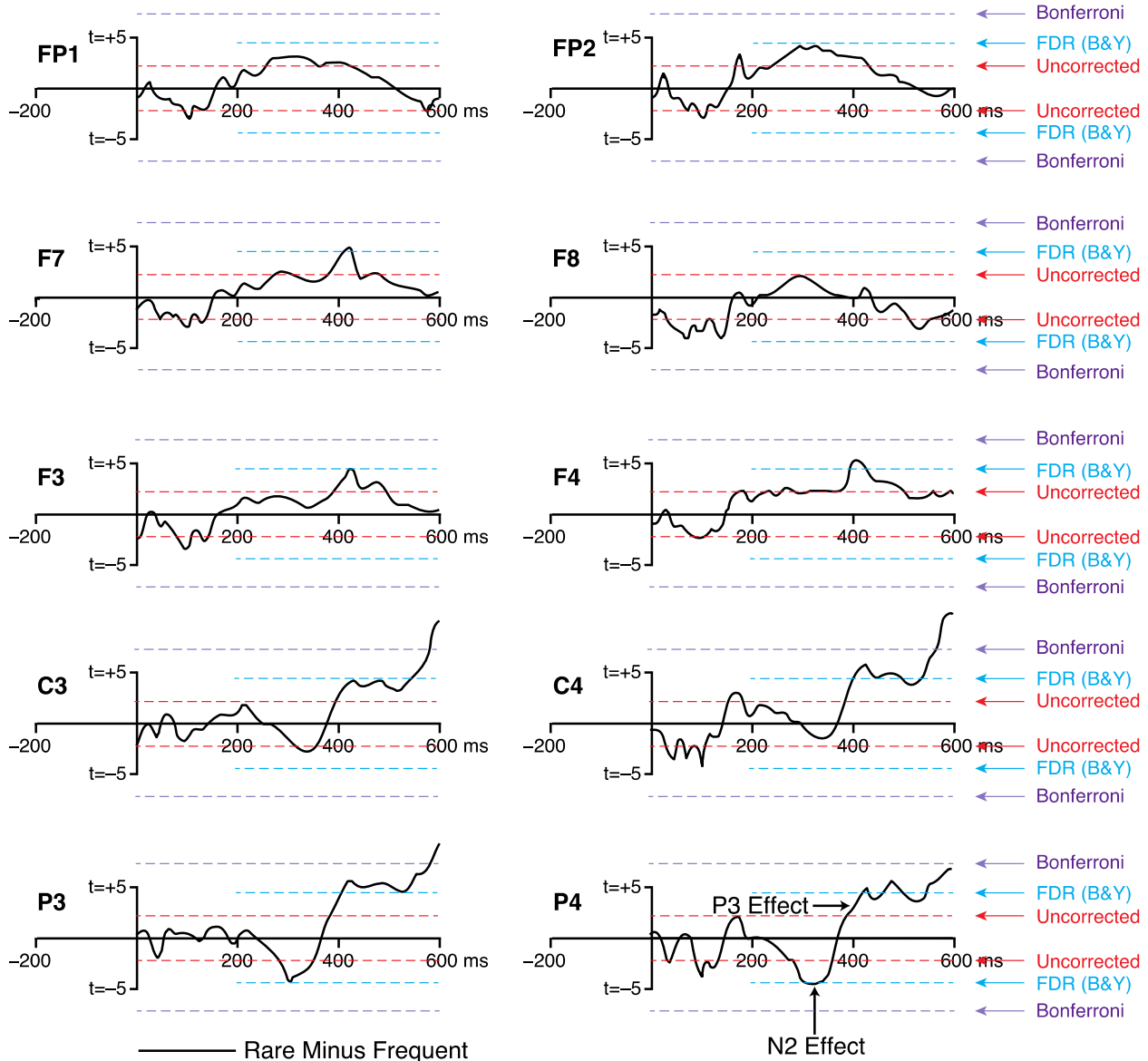


Figure 13.2. Values of the  $t$  statistic from a comparison of the rare and frequent trials in the experiment shown in Figure 13.1, computed separately at each time point for each electrode site. The data were downsampled to 125 Hz (one sample every 8 ms) before the  $t$  values were computed. The red dashed line shows the uncorrected threshold for statistical significance. The violet dashed line shows the Bonferroni-corrected threshold. The light blue dashed line shows the threshold after correction for the false discovery rate (FDR), using the formula provided by Benjamini and Yekutieli (2001) (B&Y).

Figure 13.2 shows the 3000  $t$  values from the combination of 300 time points and 10 electrode sites in our example oddball experiment. The  $t$  values are plotted like ERP waveforms, but each point in a waveform is the  $t$  value from a comparison of the rare and frequent ERPs from the 12 subjects at a particular combination of time and electrode<sup>1</sup>. Without any correction for multiple comparisons, we would treat any  $t$  value that exceeds  $\pm 2.201$  as being significant (because a  $t$  value of 2.201 corresponds to a  $p$  value of .05 with 11 degrees of freedom). This *critical value of  $t$*  is shown with red dashed lines in Figure 13.2. You can see that many of the  $t$

<sup>1</sup> You should note that the N2 effect looks almost as big as the P3 effect in the  $t$  plot, whereas the P3 effect is much larger than the N2 effect in the voltage plot. This is because the P3 effect is much more variable than the N2 effect, which counteracts the larger absolute size of the effect.

values would be significant if we used this uncorrected approach. This includes effects that are real, such as the P3 effect from approximately 400-600 ms at F3/4, C3/4, and P3/4 and the N2 effect from approximately 250-350 ms at the C3/4 and P3/4 sites (we know these are real because they've been seen in countless previous experiments). However, several bogus effects also exceed this uncorrected threshold for significance, such as the negative “blip” in the P3 and P4 electrodes at approximately 50 ms. Clearly, it would not be legitimate to use this threshold for significance.

To be considered significant with the Bonferroni correction, a  $t$  value would have to exceed  $\pm 7.24$ . Dashed lines showing this threshold are also shown in Figure 13.2. With this highly conservative threshold, the only significant  $t$  values are at the very end of the P3 wave at the C3, C4, and P3 electrode sites. Most of the time period of the P3 effect is not significant, and none of the N2 time points is significant. If we used this correction, we would falsely accept the null hypothesis for these time periods, and many experiments with real effects would not have any time points that were considered significant.

Let's think about what the Bonferroni correction tells us. Imagine that you analyzed the data from an experiment and found that 100  $t$  values were significant after the Bonferroni correction. You would have 95% confidence that *every single one* of those 100 differences was real. In other words, it would be very unlikely that even one of those 100 differences was bogus. For example, if the 100 significant  $p$  values occurred from 400-500 ms at the P3 and P4 electrode sites, you would have great confidence that each and every one of those time points differed between the rare and frequent trials. There would be only a 5% chance that even one of those significant time points was bogus. That's an extremely stringent standard.

### The False Discovery Rate (FDR) Correction

The FDR correction—which was originally developed by the statisticians Yoav Benjamini and Yosef Hochberg (Benjamini & Hochberg, 1995)—gives us a different standard for significance, one that may be more appropriate in some situations. Instead of controlling for the possibility that any single one of the significant  $t$  values is bogus, it controls for the total proportion of bogus effects. In other words, if a set of  $t$  values are significant after the FDR correction, we would expect that 5% of them are bogus. In contrast, if a set of  $t$  values are significant after Bonferroni correction, there would be only a 5% chance that one or more of them are bogus<sup>2</sup>. For example, if we find 100 significant  $t$  values with the FDR correction, we would expect that 5 of them would be bogus. However, if we find 100 significant  $t$  values with the Bonferroni correction, we would expect that none of them are bogus (because there would be only a 5% chance that even one would be bogus). Thus, the FDR correction gives us *weak control* over the Type I error rate, because it does not control the Type I error rate for every individual statistical test.

There is a very important detail that you need to know about the FDR correction. Specifically, there are different formulas for doing the correction, and they reflect different assumptions about the data (see Groppe et al., 2011a for details; and see Groppe, Urbach, & Kutas, 2011b for simulations). The most common version (described by Benjamini & Hochberg, 1995) assumes that the various values being tested are either uncorrelated or positively correlated, and it is not guaranteed when the values are negatively correlated. A negative correlation will happen if you have both positive-going and negative-going effects, either at different time points (e.g., the N2 and P3 in our example oddball experiment) or at different electrode sites (e.g., if some of your electrodes are on the opposite side of the dipole). When I applied this version of the FDR correction to the oddball data, it said that the critical value of  $t$  was 2.98, meaning that any  $t$  values that were less than -2.98 or greater than +2.98 should be considered significant. This included many effects between 0 and 150 ms that were clearly bogus. This might just reflect the 5% of bogus points that would be expected to be exceeded the

---

<sup>2</sup> This is an informal way of describing it. A  $p$  of .05 doesn't exactly mean that there is a 5% chance that the effect is bogus. Instead, it means that we would falsely reject the null hypothesis in 5% of experiment in which we used this criterion. But describing it that way requires about 20 times as many words, and I'm willing to give up a little precision to make the main principle easier to understand.

threshold for significance. Alternatively, it may be an indication that this version of the FDR correction becomes inaccurate when the data set contains effects that go in opposite directions, creating negative correlations. It is difficult to know from a single simulation like this.

Groppe et al. (2011b) did not find an inflation of the Type I error rate with this FDR approach in a set of simulations, but their simulations did not include anything like the broad N2-followed-by-P3 pattern of the present oddball experiment. More simulation studies are needed to determine the conditions under which the different FDR procedures are appropriate. If you have both positive- and negative-going effects in the same waveforms and you want a correction that is guaranteed to work accurately even with negatively correlated effects, I would recommend using the FDR correction described by Benjamini and Yekutieli (2001). However, this is a conservative correction, so it may increase your Type II error rate. It is quite possible that the original FDR correction will turn out to be just fine under most realistic conditions.

Figure 13.2 shows what happened when I used the Mass Univariate Toolbox to apply the Benjamini and Yekutieli (2001) FDR procedure to the data from our example oddball experiment. I did two “tricks” to increase the power of the analysis. First, I downsampled the data from 500 Hz to 125 Hz (one sample every 8 ms). This involved applying a simple filter and then taking every fourth data point. As a result, the procedure needed to compute only 75  $t$  values at each electrode site, for a total of 750  $t$  values instead of 3000  $t$  values. This won't change the proportion of bogus effects obtained with an FDR correction, but it reduces the raw number (and can be very useful with other correction procedures). Second, instead of testing every point between 0 and 600 ms, I tested every point from 200 to 600 ms. Prior experience tells me that it's unlikely that we will see a significant effect prior to 200 ms in an oddball experiment. Thus, I can use this a priori knowledge to reduce the number of comparisons even further (50  $t$  values per electrode, or 500 total). Reducing the time range in this manner can dramatically increase your statistical power. The resulting critical value of  $t$  was 4.41, which is shown in Figure 13.2. If we used this criterion for significance, we would conclude that the P3 effect was significant from approximately 400–600 ms at the parietal electrode sites (with some points also being significant at the central electrode sites), along with a portion of the N2 effect at the P4 electrode site. This is a big improvement over the Bonferroni correction. However, our control over the Type I error rate is now weaker, because we would expect that 5% of the significant points are actually bogus (as opposed to a 5% chance that even one significant point was bogus).

Sometimes people say that the FDR correction is less conservative than the Bonferroni correction. However, that's not exactly right. If the null hypothesis is true for all of the comparisons (e.g., if the rare and frequent trials were actually identical at all time points for all electrode sites), the probability of finding 1 or more significant differences is actually the same for FDR and Bonferroni (i.e., the Type I error rate is .05 for both corrections when the null hypothesis is true for all comparisons). Thus, if you see some significant effects with the FDR correction, you can be sure that there really is a difference at some points and at some electrode sites (with the usual 5% chance of a false positive), and you can reject the null hypothesis with the same level of confidence as you can using the Bonferroni correction. However, if there is a real effect at a large number of time points, you will typically see more significant time points with the FDR correction than with the Bonferroni correction, and more of these significant time points are likely to be bogus with the FDR correction. Thus, you can't trust every individual significant time point with the FDR correction, but you can assume that only 5% of them are likely to be bogus. This is just an average, however, and the proportion of bogus significant time points will occasionally be quite a bit greater than this (and occasionally quite a bit smaller).

In many studies, we don't care about the significance at every time point and in every channel. We just want to know if there is a difference between conditions overall, with some useful information about which time points and channels exhibited the effect. This is where the false discovery rate (FDR) correction becomes helpful.

## The $t_{\max}$ Permutation Approach

The FDR correction is a straightforward extension of the classic “parametric” approach to statistics, which makes a variety of assumptions about the nature of the data to provide computationally simple, analytic solutions. An alternative and increasingly popular class of statistical tests use the properties of the actual data to assess statistical significance rather than relying on parametric assumptions. The particular variant that I will describe here is called the *permutation* approach, because it involves examining random permutations of the data to estimate the *null distribution* (the distribution of t values that would be expected by chance if the null hypothesis were true).

### *The Basic Idea*

The most common type of permutation approach used in ERPs begins by computing a t value comparing two conditions at each time point for each electrode site, just as in the FDR approach. However, it uses a different approach for figuring out the critical value of t that is used to decide which t values are significant. Whereas the FDR approach uses a set of assumptions to reach an analytical solution to this problem, the permutation approach applies brute force computer power to determine how big a t value must be to exceed what we would expect by chance. In addition, the FDR approach provides weak control of the Type I error rate (i.e., 5% of the points identified as being significant will typically be bogus), whereas the permutation approach described here provides strong control just like the Bonferroni correction. However, the permutation approach improves upon Bonferroni correction by taking into account the fact that the individual t tests are not actually independent.

The permutation approach is related to the simulation shown in Figure 10.6 (in Chapter 10), where I took the standards from the oddball experiment and randomly divided them into two subsets, called A and B. Even though the two conditions were sampled from the same set of trials, I was able to find some time periods during which there were differences that exceeded the usual criterion for statistical significance. I did this by measuring the mean amplitude during a measurement window of 50-150 ms (which I chose by looking at the grand average waveforms). This was a very informal approach. A more formal approach would be to compute a t value comparing conditions A and B at each time point for each electrode site with these data (as in the FDR approach). Because of the large number of comparisons, this approach would also yield many time points with large t values. If we looked at the t values from these randomized data, this would give us a sense of the kinds of t values we might expect by chance in the comparison of the targets and standards if the null hypothesis were true (i.e., if targets and standards elicited identical ERPs).

One small change is necessary, however. In the simulation shown in Figure 10.6, I subdivided the standard trials into two subsets, one with 512 trials per subject and one with 128 trials per subject. However, the actual experiment contained 640 standards and 160 targets per subject, so we would need to make averages that have this number of trials to provide a fair comparison. If the null hypothesis were true in the actual experiment, then this would mean that the ERPs elicited by the standards and the targets were identical, except for noise. Thus, we can create a simulation by taking these 800 trials and randomly resorting them into one set of 640 trials and another set of 160 trials (for each subject). If we averaged the trials from within each set for each subject, we would get something that would look much like the simulation shown in Figure 10.6, but it would be based on the same trials (and same number of trials) as used in the comparison of the targets and standards in the real experiment (as shown in Figure 13.1).

If the null hypothesis is true (i.e., the ERPs are exactly the same for standards and targets), creating averages by randomly assigning stimuli to the “target” and “standard” categories should be no different from assigning the true targets to the “target” category and the true standards to the “standard” category (except for random noise). Thus, randomly resorting the event codes in this manner gives us a way of seeing what the t values should be like if the null hypothesis is true. With a few more steps, we can use this to figure out the

distribution of  $t$  values should be like if the null hypothesis is true. We can then use this to determine how large a  $t$  value must be in the analysis of the real data to be considered significant.

Once the event codes have been randomly permuted and averaged ERP waveforms have been made for the fake “target” and “standard” events, the next step is to compute a  $t$  value comparing the “target” and “standard” trials at each time point for each electrode site. Because we’ve randomly sorted the data into the “target” and “standard” waveforms in this simulation, we know that the null hypothesis is true, and any large  $t$  values must be bogus. Imagine, for example, that we looked for the largest  $t$  value across all time points and channels (which I will call  $t_{\max}$ ), and we found that this  $t_{\max}$  value was 4.2. This suggests that we might expect to find a  $t$  value of 4.2 just by chance in a comparison of the real targets and standards, even if the null hypothesis is true. Consequently, we shouldn’t take a  $t$  value seriously in a comparison of two real conditions unless it is substantially greater than 4.2. However, this is based on a purely random assignment of trials to the “standard” and “target” averages, and our random assignment may not have been representative. For example, if we did another simulation with another random sorting (permutation) of the trials, we might find a  $t_{\max}$  of 5.1. However, if we did enough of these random simulations, we could get an idea of what the  $t_{\max}$  is likely to be. For example, if we did this simulation 1000 times, with a different random sorting of the trials in each iteration of the simulation, we could get a very good estimate of the likelihood of getting a particular  $t_{\max}$  value by chance. This is exactly what we do in a permutation analysis. That is, we try many different random permutations so that we can estimate the distribution of  $t_{\max}$  for a null effect.

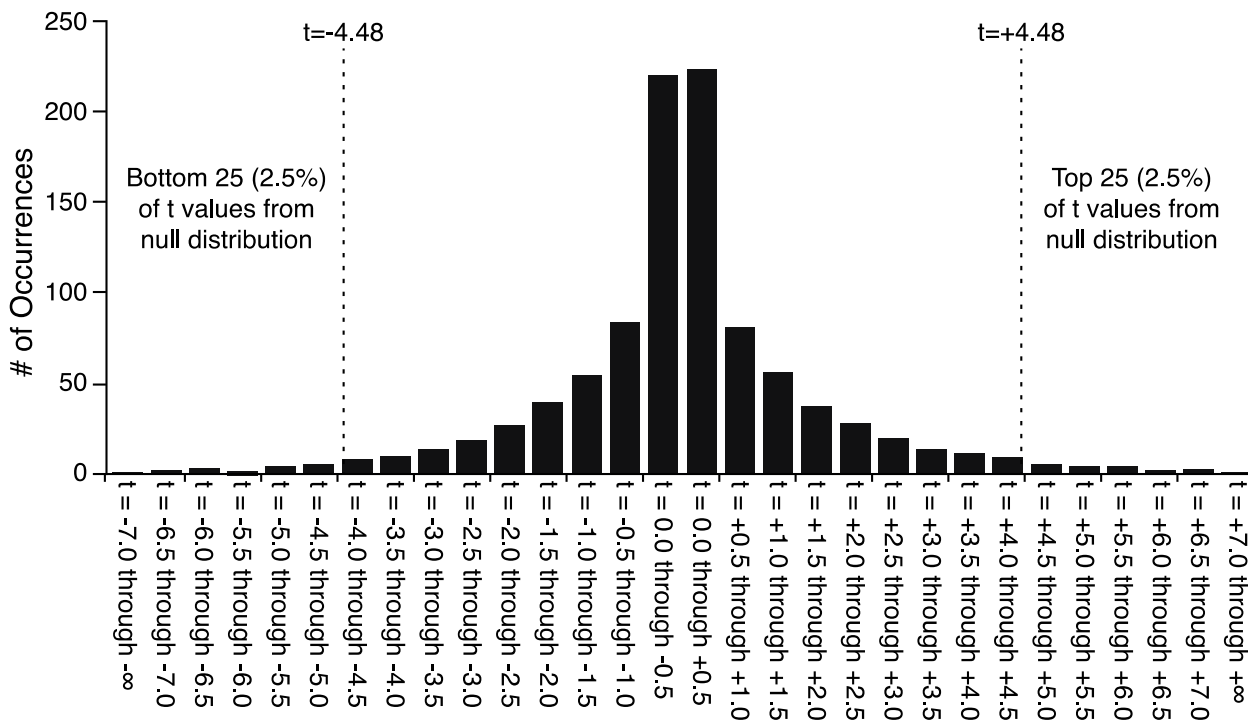


Figure 13.3. Frequency distribution of  $t$  values obtained by permuting the event codes in the experiment shown in Figure 13.1. Each bar represents the number of occurrences of a  $t$  value within a particular range over the course of 1000 iterations of the permutation procedure.

Figure 13.3 shows the distribution of  $t_{\max}$  values I obtained from this kind of random permutation procedure. On each of the 1000 iterations, I shuffled the event codes, made the averages, computed a  $t$  value comparing the “standard” and “target” waveforms across the 12 subjects for each time point at each electrode site, and found the  $t_{\max}$  value. After the 1000 iterations were complete, I counted the number of times I obtained various  $t_{\max}$  values. (My computer actually did the shuffling, averaging, and counting, of course.) For example, Figure 13.3 shows that the  $t_{\max}$  was between 0.0 and +0.5 on 224 of the 1000 iterations. The overall distribution of values is an estimate of the *null distribution* for  $t_{\max}$ . This distribution tells us how likely it is to obtain a

particular  $t_{\max}$  if the null hypothesis is true. When we look at the  $t$  values obtained from the real (non-permuted) data, we will call a time point significant if the  $t$  value for this time point is more extreme than the 5% of most extreme  $t_{\max}$  values from the null distribution ( $\pm 4.48$  in this particular example<sup>3</sup>). Thus, if we see any  $t$  values of  $< -4.48$  or  $> +4.48$  in the real data, we will consider the effects at those time points significant. Note that this is exactly how conventional statistics use the  $t$  value to determine significance, except that they use assumptions about the underlying distribution of values to analytically determine the null distribution of  $t$ . With the permutation approach, we instead determine the null distribution from the observed data. This allows us to account for things like the number of  $t$  tests being computed, the correlations among different time points, and the correlations among different electrode sites.

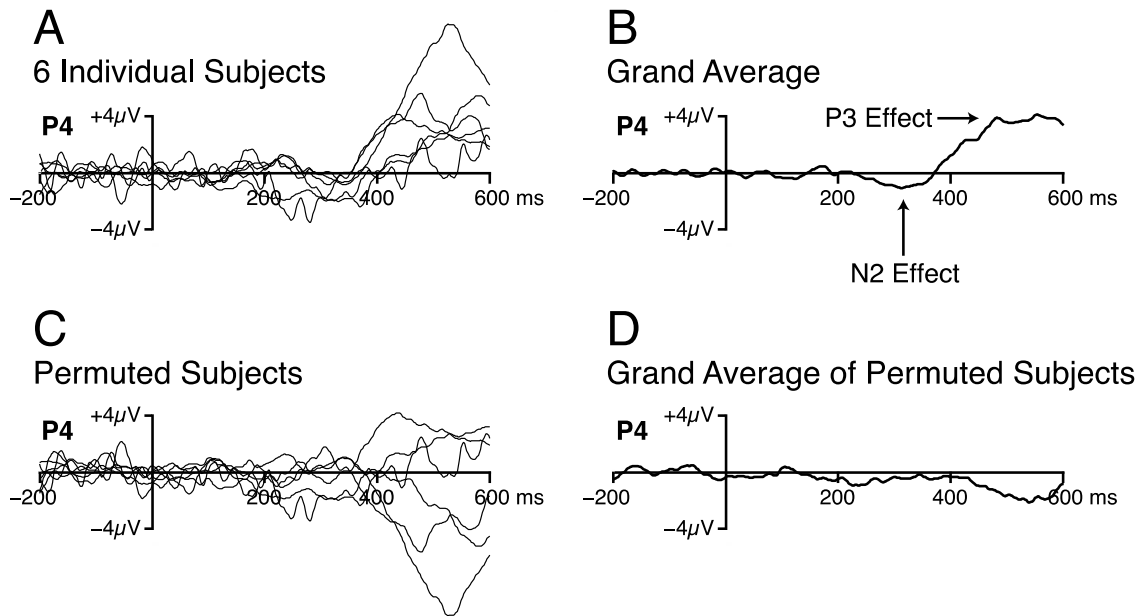


Figure 13.4. Example of the whole-subject permutation approach, showing a subset of 6 of the 12 subjects in the example oddball experiment. (A) Rare-minus-frequent difference waves from each of the 6 subjects. (B) Grand average of the difference waves from (A). (C) Permuted version of the data in (A), in which the waveforms from half of the subjects were inverted (which is equivalent to reversing the rare and frequent conditions for these waveforms). (D) Grand average of the permuted difference waves from (C).

### Making it Faster

As you can imagine, it is very time-consuming to do this shuffling and averaging 1000 times. In most cases, you can use a little trick that makes it go much faster. Rather than randomly shuffling the event codes for each subject and then averaging, you can just take a random subset of the original averaged waveforms and invert them. This is like pretending that the targets were the standards and the standards were the targets for these subjects. If the null hypothesis is true, then the averaged waveforms for two conditions will be equivalent except for noise, and it shouldn't matter if we treat the targets as standards and the standards as targets for some of the subjects. I call this a *whole-subject* permutation approach. This is what is implemented in the Mass Univariate Toolbox. It's much faster than the approach I described in the preceding section, because the data do not need to be re-averaged for each permutation.

<sup>3</sup> The boundaries I'm using ( $-4.38$  and  $+4.38$ ) give us the bottom and top 2.5% of values. Together, these top and bottom "tails" of the distribution give us the 5% of most extreme  $t$  values from the null distribution. In this experiment, a negative  $t$  value would mean that the voltage for the target was more negative than the voltage for the standard. If we had a directional effect (e.g., if we were looking only for a larger P3 for the targets than for the standards), it might be justifiable to use the most positive 5% of values from the null distribution as our criterion for significance. This would be a one-tailed test. However, this would mean that we would ignore any points with negative  $t$  values, no matter how large they were (including a larger N2 for the targets than for the standards). One-tailed tests are therefore justifiable only rarely.



Figure 13.4 shows an example of this whole-subject permutation approach from 6 subjects in our oddball experiments (the figure would be too crazy if I included all 12 subjects). To make things simple, the figure shows rare-minus-frequent difference waves (and the statistical question is therefore whether a given voltage is significantly different from zero). Panel A shows the actual difference waves from the 6 subjects, and Panel B shows the grand average of these 6 difference waves. You can see the N2 and P3 effects quite easily in this grand average. Panel C shows the waveforms after the whole-subject permutation, in which the waveforms from 3 of the 6 subjects have been inverted (which is equivalent to treating the targets as standards and the standards as targets in these subjects). Panel D shows the grand average of the permuted set of waveforms, which does not contain any large deviations from zero. If the null hypothesis is true, then the deviations from zero in the true grand average (Panel B) are just the result of chance.

In the whole-subject permutation approach, different random subsets of the averaged ERP waveforms are inverted on each iteration. The  $t$  values are then calculated at each time point and electrode site from the resulting set of averaged ERP waveforms (in which some of the waveforms have been inverted and others haven't). This is repeated 1000 times (or maybe more), which gives us the same kind of null distribution shown in Figure 13.3.

When I performed this analysis on the example oddball experiment shown in Figure 13.1, it indicated that the critical value of  $t$  was 4.49. In other words, a given  $t$  value is considered significant if it is less than -4.49 or greater than +4.49. This is quite similar to the critical  $t$  value that I obtained with the FDR approach (4.41), which means that the two approaches identified almost exactly the same points as being significant. However, the permutation results provide strong control over the Type I error rate, whereas the FDR results provide only weak control. With the permutation results, we can have high confidence in every single significant point. The FDR and permutation results will not always be this similar. I haven't applied the two approaches to a large number of different experiments, but I would expect that the FDR approach will typically identify more points as being significant when there really is an effect somewhere in the data (but more of these points are likely to be bogus).

You can also apply the permutation approach to between-subjects comparisons. Imagine, for example, that we had rare and frequent stimuli in a patient group and in a control group. You could make a rare-minus-frequent difference wave for each subject, and compute a  $t$  value comparing the two groups of subjects at each time point for each electrode site. You could then do a series of permutations, in which you randomly reassigned subjects to the "patient" and "control" groups, and calculate  $t_{\max}$  from each iteration. This would give you the null distribution for  $t_{\max}$ , and any  $t$  values in the comparison of the true groups that exceeded the 95% range of this null distribution would be considered significant differences between groups.

It's also possible to extend this to comparisons that involve three or more conditions. This is done by simply computing the  $F$  value for the relevant comparison instead of a  $t$  value. You can also do this with interactions instead of main effects. It's very versatile.

### *Does This Really Work?*

Yes, it works. It's valid and robust for the vast majority of ERP experiments. You might be concerned that the variance is greater at late time points than at early time points (or at some electrode sites relative to others). However, this doesn't distort the  $t$  values that would be obtained if the null hypothesis is true, so it doesn't bias the results in any way.

One caveat is that it that the permutation approach may be distorted by differences in variance between the populations when you conduct between-group analyses (e.g., if a patient group exhibits more variability than a control group). This can be particularly problematic when the sample sizes also differ across groups. There are ways of dealing with this, but you will need to be careful in this situation (see Groppe et al., 2011b for details). For within-subjects designs, however, all the evidence indicates that the permutation works appropriately under

a wide range of realistic conditions. Indeed, permutation statistics should work under a broader range of conditions than conventional parametric statistics because they make fewer assumptions.

### **Looking for Clusters**

In fMRI experiments, real effects are typically present across many contiguous voxels, and statistical power can be gained by finding multiple contiguous voxels in which each voxel is significant at some uncorrected alpha level (e.g.,  $p < .01$ , uncorrected). Permutation statistics are then used to determine the number of contiguous voxels that would be expected by chance (the null distribution), and a cluster is deemed significant if the number of voxels is greater than would be expected by chance (greater than 95% of the clusters found when the data have been permuted many times).

The same approach can be used with ERPs, under the assumption that real effects will typically occur over multiple consecutive time points and in multiple adjacent electrode sites (Groppe et al., 2011b). To make this easier to understand, let's consider how I applied this approach to our example oddball experiment. Figure 13.5 shows the end result, with yellow bars indicating time points that were significant in a *cluster mass* analysis. To perform this analysis, I first resampled the data to 125 Hz to reduce the number of points to be tested (as I did in the FDR analysis). I then instructed the Mass Univariate Toolbox to look for clusters of significant differences between the rare and frequent trials between 200 and 600 ms.

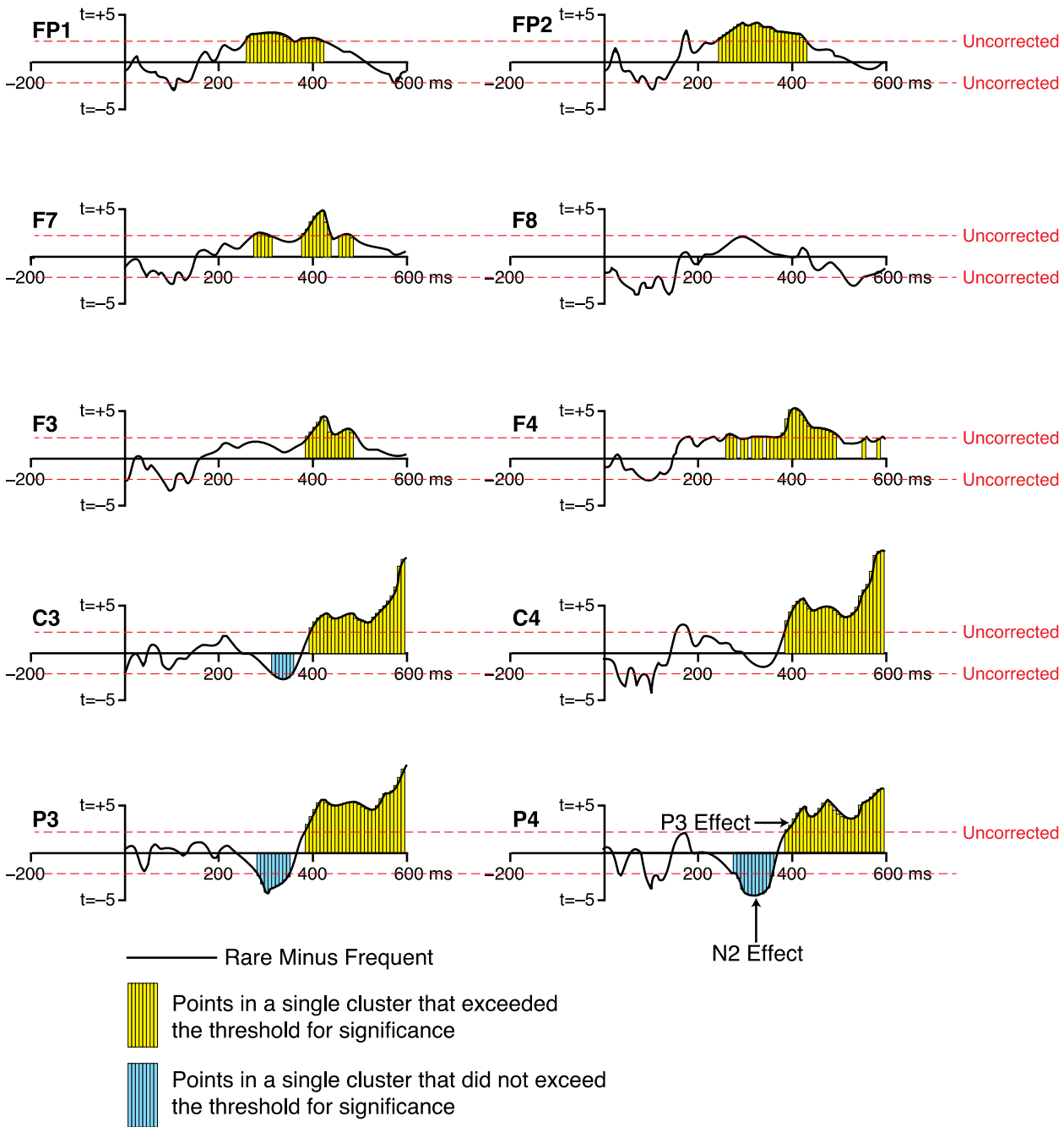


Figure 13.5. Results of the cluster mass analysis of the data shown in Figure 13.1. As in Figure 13.2, the waveforms show the  $t$  values from a comparison of the rare and frequent trials, computed separately at each time point for each electrode site, and the uncorrected threshold for significance is indicated by the red dashed line. Each yellow rectangle represents a  $t$  value that was part of a single significant cluster, which extended over many time points and almost all of the electrode sites. Each blue rectangle represents a  $t$  value that was assigned to a single cluster that did not meet the threshold for significance (although each point in the cluster exceeded the uncorrected threshold for significance).

The algorithm began by computing  $t$  values at each time point for each electrode site, as in the FDR and permutation analyses. The red line in Figure 13.5 shows the uncorrected threshold for significance ( $p < .05$ ). The algorithm looked for clusters of  $t$  values in which each  $t$  value in the cluster exceeded this threshold. A point with a significant  $t$  value was assigned to a particular cluster if it was temporally adjacent to another point with a significant  $t$  value at the same electrode site or if it was within a particular spatial distance of another significant value at the same time point. For example, there was a significant  $t$  value at 400 ms at the Fp1

electrode site, and there was also a significant  $t$  value at 408 ms, so these two  $t$  values were assigned to the same cluster (note that we have a sample every 8 ms given that we resampled the data at 125 Hz, so 400 ms is temporally adjacent to 408 ms). Similarly, there was also a significant  $t$  value at 400 ms in the F7 electrode site, so this was assigned to the same cluster (because F7 is close to Fp1). The clustering process finds all clusters of 2 or more temporally or spatially adjacent significant  $t$  values (creating clusters that are as large as possible). Note that the  $t$  values within a cluster are required to have the same sign (because a significant negative value followed immediately by a significant positive value is unlikely to reflect a single underlying effect).

The algorithm then sums together all the  $t$  values within a given cluster; this sum is the *mass* of the cluster. The idea is that a large cluster mass is likely to happen with a real ERP effect (because ERP effects tend to be extended over multiple time points and over multiple electrode sites). However, a large cluster mass is unlikely to happen with noise (although some clustering will happen because noise may also be distributed over adjacent time points and nearby scalp sites to some extent). To determine whether the mass of a given cluster is larger than what would be expected from chance, a permutation analysis is conducted. The conditions are shuffled many times (usually using the whole-subject approach), and for each iteration the largest cluster mass is found in the permuted data. The null distribution of cluster masses is then determined from the multiple iterations, and the 95% point in this null distribution is determined. Cluster masses from the analysis of the original data are considered significant if they are greater than the 95% point in the null distribution.

When I applied this approach to the data from the example oddball experiment, the algorithm found 2 clusters of positive  $t$  values and 3 clusters of negative  $t$  values. Only one of these 5 clusters had a mass that was beyond the 95% point in the null distribution, so this was the only cluster that was considered significant. All of the values marked as significant with yellow vertical bars in Figure 13.5 belong to this one cluster. If you look carefully, each significant value is temporally and/or spatially adjacent to another significant value. This is basically a P3 effect that is significant over a broad range of time points and a broad range of electrodes. Note that there may be gaps between significant points, both in time (as in the gap between ~325 and 375 ms at F7) and in space (as in the gap in the F8 electrode site between the Fp2 and F4 sites). These are likely to be Type II errors, and they do not mean that the effect has truly gone away in the gaps.

The light blue bars in Figure 13.5 show a nonsignificant negative cluster from approximately 250-350 ms at the C3, P3, and P4 electrode sites. That is, there are several consecutive points within this time period that go beyond the uncorrected threshold for significance, and these points are at nearby electrode sites. This corresponds to the N2 component, which really is larger for the rare stimuli than for the frequent stimuli (as we know from many prior oddball experiments). However, it was not significant in this cluster mass analysis. The mass of the cluster formed by this N2 effect was simply not large enough to go beyond what would be expected by chance. Indeed, you can see a bogus cluster of individually significant  $t$  values from approximately 50-150 ms at many of the frontal electrode sites.

One important detail to note about cluster-based approaches is that electrodes at the edge of the electrode array will have neighbors on only one side, so this means that they have fewer neighbors<sup>4</sup>. This will tend to reduce the statistical power a bit at these electrodes. Also, cluster-based approaches provide only weak control over the Type I error rate (i.e., you can't have confidence in each and every point that was classified as significant).

## Practical Considerations

There are two important practical considerations that you should keep in mind with the mass univariate approach. The first is to decide which approach to take (e.g., individual-point FDR, individual-point permutation, cluster-based permutation). The second is to maximize your statistical power by adding in as many a priori constraints as possible.

---

<sup>4</sup> This was explained to me by David Groppe, who in turn heard it from Manish Saggar.

### *Which Approach is Best?*

First, let's compare FDR-based versus permutation-based methods for controlling the Type I error rate when the individual points are being assessed. If you mainly want to know if the conditions differ at all (i.e., whether or not the null hypothesis is true at all time points for all electrode sites), either approach should be fine. As I mentioned previously, the FDR approach is just as conservative as the Bonferroni approach when the null hypothesis is actually true. That is, if you get at least one significant point, you can be confident that there is a real effect somewhere in the data. The permutation approach also guarantees strong control over the Type I error rate when the null hypothesis is true for all points. However, if you have good reason to believe that an effect is present, and you are mainly interested in determining the time range and scalp distribution of the effect, the FDR and permutation approaches will typically yield different results. If you are willing to tolerate the idea that approximately 5% of the points identified as being significant are actually bogus, the FDR approach is fine. However, if you want to have confidence in every single significant point, you should use the permutation approach. For example, if you find that the earliest significant effect is at 158 ms, and you want to draw a strong conclusion based on the latency of this effect, you should use the permutation approach (because it provides strong confidence in every significant point). However, if you are just trying to describe the general time course of an effect, without emphasizing any particular time point, the FDR approach may give you greater power.

Now let's consider whether you should use a point-based approach or a cluster-based approach. In our example oddball experiment, the cluster-based analysis did not identify the same significant points as the point-based FDR analysis (or the point-based permutation analysis, which was nearly identical to the point-based FDR analysis). Specifically, the P3 effect was significant over a broader range of time points and electrode sites in the cluster-based analysis than in the point-based FDR analysis, whereas the N2 effect was significant (although barely) at the P4 electrode in the point-based analysis but not in the cluster-based analysis. Which analysis is giving us the correct answer? The reality is that statistical analyses are not designed to produce an absolute "correct" answer. They are designed to give us a standard for deciding which effects to believe, while controlling the probability that we will be incorrect. In general, our beliefs are (and should be) shaped by both the current data and prior knowledge. If our prior knowledge tells us that an effect is likely to be distributed over a broad range of time points and electrode sites, then the cluster-based analysis will give us good power to find such an effect if it is present (without biasing us to find an effect of this nature if it is absent). If we don't want to make this assumption, the FDR-based single-point approach will give us more power to find effects that are more tightly focused in time or scalp distribution (such as the N2 effect in our example experiment). In general, I would recommend using the cluster-based approach only when you are interested in broad effects.

Could you use multiple approaches for a given experiment? Unfortunately, the answer is almost always no. This would be "double dipping," and it would inflate your Type I error rate.

### *Maximizing Power*

If you test for significance at 500+ time points for 32+ electrode sites with an FDR- or permutation-based correction for multiple comparisons, you will have very low power for detecting significant effects. This is because you will be computing an enormous number of t values, which increases the possibility that a large t value can arise by chance. If you can apply prior knowledge to reduce the number of t values being computed, you will almost always improve your statistical power. For example, when I applied the FDR correction to the data from the oddball experiment, I downsampled the data to 125 Hz (which reduced the number of t tests by 75%) and I limited the time range to 200-600 ms (which eliminated the t tests between 0 and 200 ms). In general, I recommend downsampling to the lowest rate that still gives you meaningful temporal resolution (approximately 100 Hz is usually reasonable) and limiting the analysis period to the shortest period and smallest number of electrodes that can be justified on the basis of prior knowledge. In addition, moderate low-pass filtering can also be helpful (e.g., a half-amplitude cutoff at 30 Hz), because it reduces the subject-to-subject

variance at any given time point and thereby reduces the number of high  $t$  values that you would obtain by chance.

As an example of using prior knowledge, recall that the N2 effect in the oddball experiment was not significant in the cluster-based analysis. Because previous research tells me that this effect is always between 200 and 400 ms and has a posterior scalp distribution, it would be justifiable to limit the analysis of the present oddball experiment to a 200-400 ms time window and the P3 and P4 electrodes. When I did this, I found a significant cluster at that included both P3 and P4 and extended from approximately 280-360 ms.

Although this gave me more power to detect the N2 effect, I want to make it clear that a priori knowledge can be a double-edged sword. On one hand, constraining the time points and electrodes can increase your power to detect effects in the expected time windows and electrode sites. On the other hand, if you constrain your analyses in this way, and you get a large but unexpected effect in some other time window or electrode sites, you are not allowed to change your time window or sites to include this effect. In an ideal world, you would decide on the time window and electrode sites before seeing the data. If you start making choices about time windows and electrode sites after you've seen the data, your choices will inevitably end up increasing your Type I error rate. In the real world, however, it is extremely difficult to make all of these decisions before you see the data. Moreover, even if you intended to focus on 300-800 ms at the posterior electrode sites, are you really going to ignore a very large and interesting frontal effect at 250 ms?

This brings us back to the most important principle for assessing statistical significance, which was raised in Box 10.1 in Chapter 10: Replication is the best statistic. If you conduct your statistical analyses using minimal a priori assumptions, you will have low power for detecting real effects. You may therefore add some assumptions to increase your power, but these assumptions will probably be biased by the observed effects, increasing the probability of a Type I error. Thus, you won't have real confidence in your effects until you replicate them.

The bottom line is that you should do whatever you can within reason to minimize inflation of the Type I error rate when analyzing a particular study, and then you should replicate your findings before you fully trust them. And if you are evaluating someone else's research, you should make sure that the analysis approach does not inflate the Type I error rate, and you should ask to see replications for effects that are surprising.

## References

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165-1188.
- Maris, E. (2004). Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology*, 41, 142-151.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164, 177-190.
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology*, 48, 1711-1725.
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology*, 48, 1726-1737.
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*.
- Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology*, 49, 549-565.