A Bayesian Approach to the Correction for Multiplicity

Tim de Jong
Thesis for the Research Master in Psychology, University of Amsterdam
Advisors: Maarten Marsman and Eric-Jan Wagenmakers

Abstract

There are many situations in which researchers perform multiple hypothesis tests simultaneously. It is important that the results of these tests are corrected for multiplicity. If this correction is not performed, it is likely that some null hypotheses will be falsely rejected. There are various different methods for performing multiplicity corrections, dependent on the specific type of multiple testing. If you find yourself in the frequentist camp and wish to conduct pairwise comparisons following a one-way ANOVA you are in luck, as methods to do so are readily available to researchers. On the other hand, a Bayesian is hard-pressed to find an appropriate correction method in this case. In this thesis we evaluate two Bayesian methods that allow pairwise comparisons while protecting against false positive results. We demonstrate the importance of dealing with the dependence structure that exists among pairwise comparisons. To aid researchers with their statistical inference our aim is to implement these methods in the statistics software JASP.

Keywords: Bayesian inference, multiplicity, pairwise comparisons, dependency

The Salmo salar — more commonly referred to as the Atlantic salmon — is a popular target for recreational and commercial fishermen. A little known fact is that these fish are surprisingly diverse; not only can they navigate to their place of birth from thousands of kilometers away, they can even perform psychological tasks designed for humans. Researchers discovered this when they placed an Atlantic salmon in an fMRI scanner and recorded brain activation as the salmon performed a mentalizing task (Bennett, Baird, Miller, & Wolford, 2009). The salmon had to view photographs of human individuals and then determine which emotion someone was experiencing. Several active voxel clusters were observed in the salmon's brain, showing task engagement. Particularly salient was the fact that the

salmon they used was not alive at the time of scanning. The purpose of this study was not to investigate the Atlantic salmon or show that fMRI studies are inherently flawed. Rather, Bennett et al. (2009) wanted to demonstrate the dangers of ignoring multiple comparisons. In a typical fMRI study over 100,000 pairwise comparisons are performed to determine which brain areas are activated. Some of these tests will yield a result purely by chance. This problem is not unique to fMRI studies; it can occur any time multiple inferences are made on a dataset. Consequently, it is not bound to just one research field; the problem is encountered in various different areas of science. In genetics ofttimes hundreds of thousands of tests are performed to determine associations between genotype and phenotype (e.g., Storey & Tibshirani, 2003). In economics, researchers are faced with finding the best trading strategy out of a large number of options (e.g., Romano & Wolf, 2005). While any clinical trial run in the field of medicine often investigates various different treatments to determine which has the greatest efficacy (e.g., Fleming, 1982).

To illustrate one does not need to perform thousands of tests for multiplicity to become an issue, suppose we have 20 hypotheses we wish to test simultaneously. In frequentist statistics we would set some significance threshold α — usually to .05. The probability of observing at least one significant result purely by chance would then be $1-(1-.05)^{20} \approx .64$; quite a bit higher than what was originally intended. This probability of detecting an effect that is not present is usually referred to as the probability of committing a Type-I error. Directly related to this is the Type-II error, which is the probability of not being able to detect an effect when it is in fact present.

Dealing with multiplicity. There are different methods of dealing with the multiplicity problem. Researchers can find a plethora of options in the frequentist literature. The main differences reside in the type of error rate one wishes to control. The two most common error rates will be discussed with reference to Table 1. Firstly, there is the Family-Wise Error Rate (FWER), which is the probability of at least one Type-I error in the family of comparisons — $p(V \geq 1)$ (Hochberg & Tamhane, 1987). This error-rate is controlled by the popular, but conservative Bonferroni procedure (Bonferroni, 1936). The procedure is conservative because of its very strict criterion and the fact that dependency between hypotheses is ignored. A second error-rate, more liberal than the FWER is the False Discovery Rate (FDR); this error rate was popularized by Benjamini and Hochberg (1995). It is defined as $\mathrm{E}(\frac{V}{R})$, or the proportion of erroneously rejected null hypotheses; this proportion is controlled by the Hochberg-Benjamini procedure. Various other error rates and procedures to control them exist, but they are beyond the scope of this thesis (see e.g., Shaffer, 1995). Generally, the frequentist post-hoc solutions share the same goal: to identify pairwise differences between sample means while guarding against falsely rejecting the null hypothesis (Maxwell, 1980).

In Bayesian statistics, explicit multiplicity control is not always needed. If a researcher has specific expectations about the plausibility of each hypothesis under consideration, it is possible to express these expectations directly in the prior model probabilities. This subjective assignment of probability results in automatic multiplicity control as noise driven, random coincidences are unlikely to get a high prior probability. However, this situation is entirely reliant on prior knowledge of the researcher and quickly becomes unmanageable as more hypotheses need to be taken into account. For example, if an experimental study examines how well ten treatments work, a total of $\binom{10}{2} = 45$ different pairwise comparisons can be made with their own hypotheses. It is clear that the number of hypotheses can quickly explode and for such situations a different approach must be taken. We can generally distinguish two methods, hierarchical models and objective adjustment of prior model probabilities; they are discussed in the next sections.

Hierarchical models. Gelman, Hill, and Yajima (2012) argued that the correction for multiplicity is inherent to Bayesian hierarchical analyses. Hierarchical models guard against multiplicity through shrinkage of the estimates towards the group mean (Kruschke & Liddell, 2017). However, performance of this approach is heavily dependent on sample size and variance in the samples (Gelman & Loken, 2013). Another issue is the fact that an estimated model in itself does not translate directly to a test of hypotheses.

Objective adjustment of prior model probabilities. As previously noted, subjectively setting individual prior model probabilities controls for multiplicity. This approach can be contrasted with an objective assignment of model probabilities, which assumes no specific information about the individual hypotheses. It is important to note that in order to correct for multiplicity, the assignment of the probabilities must depend on the number of hypotheses m. For example, say we are interested in a variable inclusion problem and

Table 1
Possible situations encountered when performing m hypothesis tests (Benjamini & Hochberg, 1995). Rows show the true situation and columns what hypothesis tests indicated. In m comparisons there are a maximum of m_0 true alternative hypotheses. The number of true null hypotheses is then simply what remains: $m - m_0$. Hypothesis tests can give false positive results denoted as S and false negative results denoted as U.

	Declared non-true signal	Declared true signal	Total
True signal	U	V	m_0
Non-true signal	T	S	$m - m_0$
Total	$m - R$	R	m

assign prior probability of $p(\mathcal{H}_0) = \xi_0 = \frac{1}{2}$ to each hypothesis $\mu_i = 0$, where $i \in \{1, ..., m\}$. For simplicity's sake these hypotheses are assumed to be independent. We can compute the a priori expected model size and its standard deviation with the moments of a Bernoulli random variable. As each inclusion is independent we can simply sum over the inclusions to obtain the expected size: $\sum_i^m \xi_0 = \frac{m}{2}$ with a standard deviation of $\sqrt{\sum_i^m \xi_0(1-\xi_0)} = \sqrt{\frac{m}{4}}$. Note that as additional noise is added, the standard deviation does not grow proportional with m. Consequently, the expected proportion of included μ's becomes tightly coupled around $\frac{1}{2}$. It is clear that no multiplicity control is provided by these prior probabilities. Rather, assignment of the prior model probabilities should be considered with regard to the entire model space. We will now turn to three different methods for objective adjustment.

Null control. The first method is based on keeping the total prior probability of finding no difference in a set of k comparisons equal to 0.5 (Jeffreys, 1938). In this approach the prior model probability that is assigned to the null hypotheses can be obtained by solving $(\xi_0)^k = \frac{1}{2}$, which is $\xi_0 = 2^{-\frac{1}{k}}$. This method was employed by Williams, Heathcote, Nesbitt, and Eidels (2016) to perform multiple post-hoc comparisons after a one-way ANOVA. A connection can be drawn to the frequentist Bonferroni correction — which also places one penalty on the family of hypotheses; in the case of k pairwise comparisons the p-value would have to exceed the inequality $p < \frac{\alpha}{k}$ to be deemed significant. Westfall, Johnson, and Utts (1997) showed that the Bayesian and frequentist approaches are similar; in fact they noted that when alternative hypotheses are likely to be true and the number of hypotheses is large, the Bonferroni's adjusted p-value and Bayesian posterior probabilities are both proportional to their respective unadjusted value multiplied by a constant dependent on k.

Single proportion. The second method relates to the situation where you have a specific belief about the ratio of signal to noise in your data. This is somewhat similar to the fully subjective Bayesian approach, however you specify your belief in an overall proportion of signal to noise, rather than assigning belief to each individual hypothesis. This approach was employed by Stephens and Balding (2009) to correct for the many comparisons performed in genetic studies. They set the prior model probabilities of finding an effect to 10^{-4} reflecting earlier research on the topic. This method seems to share similarities with the FDR insomuch they both depend on the proportion of tests that are null, but not on the number of tests itself.

Distribution of proportions. The final method assumes a distribution on the proportion with a single hyperparameter controlling all comparisons. Consequently, the separate hypotheses are no longer independent from one another. This method has been applied to the problem of variable inclusion and edge inclusion in graphs, in genetic studies and regression models, as well as multiple comparisons between means (Bogdan, Ghosh, & Tokdar, 2008; Carvalho & Scott, 2009; Q. Li & Shang, 2015; Mitra, Mueller, & Ji, 2017; Scott &

Berger, 2006, 2010). Scott and Berger (2006) showed that as an increasing number of noise coefficients were taken into account, the posterior probabilities were shrunk towards zero. This penalty was more severe for signal coefficients which lay close to zero than those that were more clearly different from zero. As such the distribution placed on the proportion automatically corrects for multiplicity. This method will henceforth be denoted as SB.

Accessibility of Bayesian multiplicity control. Even though there are several ways of dealing with the multiplicity problem in Bayesian statistics, these methods are not readily available to researchers. This is not surprising as Bayesian statistics are playing catch-up with its frequentist counterpart, which — due to its popularity among researchers — is blessed with many R packages and graphical software solutions. Recent developments have attempted to close this gap; among these attempts we find noteworthy software such as the R packages BAS (Clyde, 2017) and BayesFactor (Morey & Rouder, 2015) and the free statistical software JASP (JASP Team, 2017). As a result, certain Bayesian techniques have become much more accessible (e.g., ANOVA, t-test, regression); however, much functionality is still lacking. Even a simple one-way ANOVA cannot be further scrutinized using standard post-hoc analyses. In this thesis we intend to address that problem. Specifically, we are interested in evaluating the null control and SB methods for pairwise comparisons; the single distribution method requires too much prior knowledge to be widely applicable. Furthermore, we seek to gain more insight in how the different methods relate to each other and to the frequentist solutions. We set out to implement the two methods for use as post-hoc analyses in JASP.

This thesis is outlined as follows, we will first go over some preliminaries: definition of the models we use with their marginal likelihoods and priors. We turn to a brief discussion of pairwise comparisons and its associated problems. The first method we then discuss is null control, followed by the SB method; for both methods we show a practical example of their implementation.

Preliminaries

Given a set of m groups we can perform a total of $k = \binom{m}{2}$ pairwise comparisons, where k is the binomial coefficient. So when we have $m = 5$ groups there are $k = \binom{5}{2} = 10$ possible pairs. A common hypothesis test for pairwise comparisons is the t-test. Jeffreys (1948) proposed a Bayes factor equivalent to the frequentist version of the t-test. The Bayes factor contrasts the marginal likelihoods of the data under the null hypothesis $\mathcal{H}_0$ and its alternative $\mathcal{H}_1$. To compare two groups of observations $\mathbf{x}$ and $\mathbf{z}$ it is assumed that the

observations can be modeled as

$$x_i \sim \text{Normal}\left(\mu - \frac{\omega}{2}, \sigma^2\right), \tag{1}$$

$$z_j \sim \text{Normal}\left(\mu + \frac{\omega}{2}, \sigma^2\right), \tag{2}$$

for $i = 1, ..., n_x$ and $j = 1, ..., n_z$. Here μ is the grand mean and ω is the effect. $\mathcal{H}_0$ specifies that $\omega = 0$, while $\mathcal{H}_1$ allows it to vary. Generally, ω is reparameterized in terms of effect size $d = \frac{\omega}{\sigma}$. This reparameterization makes the effect of interest dimensionless and as a result the Bayes factor is the same regardless of the unit of measurement (e.g., milligram or kilogram). Considering d is fixed to zero in $\mathcal{H}_0$, the priors are slightly different between the models. For $\pi(\theta \mid \mathcal{H}_0)$ we use,

$$\mu = 1, \tag{3}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}, \tag{4}$$

and the prior $\pi(\theta \mid \mathcal{H}_1)$ is specified as

$$\tag{5}$$

$$\mu = 1, \tag{6}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}, \tag{7}$$

$$d \sim \text{Normal}(0, \sigma_d^2), \tag{8}$$

$$\sigma_d^2 \sim \text{inverse chi-squared}(1). \tag{9}$$

Liang et al. (2008) noted that integrating out the variance in a design where the effect size is normal and the variance an inverse chi-square distribution is equivalent to stating:

$$d \sim \text{Cauchy}. \tag{10}$$

The marginal likelihoods in the Bayes factor are defined as:

$$p(\mathbf{y} \mid \mathcal{H}_i) = \int f(\mathbf{y} \mid \theta, \mathcal{H}_i)\pi(\theta \mid \mathcal{H}_i)\, \mathrm{d}\theta, \tag{11}$$

where $\mathbf{y}$ denotes the data and $f(\mathbf{y} \mid \theta, \mathcal{H}_i)$ is the likelihood function. The marginal likelihood for each hypothesis combined with the prior model probabilities translates to the posterior

odds:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \mathbf{y})}{p(\mathcal{H}_1 \mid \mathbf{y})}}_{\substack{\text{posterior}\\ \text{odds}}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\substack{\text{prior}\\ \text{odds}}} \times \underbrace{\frac{p(\mathbf{y} \mid \mathcal{H}_0)}{p(\mathbf{y} \mid \mathcal{H}_1)}}_{\substack{\text{Bayes}\\ \text{factor}}}. \tag{12}$$

Note that if $\mathcal{H}_0$ is specified in the numerator, posterior odds greater than 1 indicate preference for the null hypothesis.

For a more elaborate discussion of the model and its priors we refer the interested reader to Ly, Verhagen, and Wagenmakers (2016) and Rouder, Speckman, Sun, Morey, and Iverson (2009).

Pairwise comparisons and dependency

Consider the simplest case possible for multiple comparisons — $m = 3$ and $k = \binom{3}{2} = 3$ — and label the three groups A, B and C. We quickly find that these comparisons are not independent from one another. To exemplify this, assume we discovered that A and B are equal, while A and C were not. Consequently, without performing any more tests, we may conclude that the means of B and C must also be unequal. Now, when we perform a correction for the number of tests and we assume these tests are independent it means we are being overly conservative. In the case of $k = 3$ it would make more sense to correct for two tests, as the last one is not an independent inference on the data, but in essence it is given for free. This is one reason that corrections such as the Bonferroni are overly conservative; while they maintain the Type-I error rate at some specified threshold α (e.g., .05), the Type-II error rate rises quickly. Additionally, the correction may be too conservative when only a few null hypotheses are true, in this specific case the actual FWER may be less than α (Westfall, 1997).

To increase the power of post-hoc tests, the following frequentist procedures have been proposed: (1) single step correction procedures such as Tukey's range test, which is similar to a t-test except uses information from all tests to calculate a q-statistic and so accounts for dependency while controlling the FWER (Tukey, 1949); (2) multi-step procedures such as Holm-Bonferroni method (Holm, 1979), which sort the p-values and then compare each to a decreasing α based on the numbers of steps taken, instead of comparing every p-value against one threshold corrected for k. Note that some Bayesian equivalents to these multi-step procedures have been proposed, with the aim to narrow the rapidly expanding model search — for example in linear regression — often encountered (Abramovich & Angelini, 2006; Chen & Sarkar, 2004).

Null control

With the marginal likelihood defined and the dependency problem in the back of our mind, we now turn to the prior model probabilities. Recall that the method of null control is based on keeping the total prior probability of finding no differences in a set of k comparisons equal to $\frac{1}{2}$. To illustrate why this is desirable, suppose we have $k = 6$ comparisons. If we assigned each null hypothesis a prior model probability of $\frac{1}{2}$, then a priori we state that the chance of finding no effect in all comparisons is $\left(\frac{1}{2}\right)^6 = .016$ which might be much lower than what we really believe. This probability drops quickly when more noise variables are added to the mix.

The implementation of Jeffreys

There are two implementations of null control we may consider. The one proposed by Jeffreys (1938) is derived as follows. Suppose we have k alternative hypotheses each with prior probability ξ_A of being true and let's call the disjunction of these hypotheses $\mathcal{H}_A$. We assign equal probability to event $\mathcal{H}_A$ and the event that $\mathcal{H}_A$ does not occur — $\mathcal{H}_0$:

$$p(\mathcal{H}_0) = p(\mathcal{H}_A) = \frac{1}{2}. \tag{13}$$

The probability of all individual alternative hypotheses $\mathcal{H}_{Ai}$ being false, is $(1 - \xi_A)^k$. And so, as this coincides with $p(\mathcal{H}_0)$ we find the inequality

$$(1 - \xi_A)^k = \frac{1}{2}, \tag{14}$$

$$\xi_A = 1 - 0.5^{\frac{1}{k}}. \tag{15}$$

And so for any $p(\mathcal{H}_{Ai})$ we find

$$p(\mathcal{H}_{Ai}) = \xi_A \tag{16}$$

$$= 1 - 0.5^{\frac{1}{k}}. \tag{17}$$

Equally, $p(\mathcal{H}_{0i})$ is now easy to compute:

$$p(\mathcal{H}_{0i}) = 1 - \xi_A \tag{18}$$

$$= 0.5^{\frac{1}{k}}. \tag{19}$$

The problem with this approach is much the same as for Bonferroni's procedure. It does not take into account the dependency between the tests and would be too conservative.

The implementation of Westfall

The second approach to null control was established by Westfall et al. (1997). They extended Jeffreys method to account for dependency by shifting the problem from the k comparisons to the m underlying groups and their means. This shift was based on the following notion. Consider a model with a grand mean μ wherein each μ_i of the m different groups can be:

$$\mu_i = \begin{cases} \mu, & \text{with probability } \tau, \\ \sim G, & \text{with probability } 1-\tau, \end{cases} \tag{20}$$

where G is some continuous distribution — which is important as it follows that the μ_i's from G can never exactly equal each other. The probability that some μ_j and μ_l are equal to μ, or $p(\mathcal{H}_{0jl})$, is simply τ^2. We can extend this to $p(\mathcal{H}_0) = p(\text{all } \mu_i = \mu) = \tau^m$. Solving for τ we get:

$$p(\mathcal{H}_0) = \tau^m, \tag{21}$$

$$\tau = p(\mathcal{H}_0)^{\frac{1}{m}}, \tag{22}$$

and when we now substitute τ in the null hypothesis of a single comparison

$$p(\mathcal{H}_{0i}) = \tau^2 \tag{23}$$

$$= (p(\mathcal{H}_0)^{\frac{1}{m}})^2 \tag{24}$$

$$= p(\mathcal{H}_0)^{\frac{2}{m}}. \tag{25}$$

As an example consider $m = 4$ and consequently $k = \binom{4}{2} = 6$. If we use $p(\mathcal{H}_0) = 0.5$, then with Jeffreys we find $p(\mathcal{H}_{0i}) = 0.5^{\frac{1}{6}} = 0.891$ and with Westfall we find $p(\mathcal{H}_{0i}) = 0.5^{\frac{2}{4}} = 0.707$. Both methods default back to $p(\mathcal{H}_{0i}) = \frac{1}{2}$ when $m = 2$ (and $k = 1$). In Figure 1 a comparison is plotted between the methods of Jeffreys and Westfall. It is clear that ignoring dependency results in a more conservative procedure.

A simulation

What remains of interest is to evaluate the performance of Westfall's implementation of null control. As we would like to draw a link between null control and the frequentist Bonferroni correction, we need a way to determine false positives for both tests. The problem is that there is no all-or-none significance testing in Bayesian statistics. There is only the continuous evidence of the alternative hypothesis over the null. Some guidelines as to what constitutes evidence for a hypothesis have been proposed (see Table 2). However,

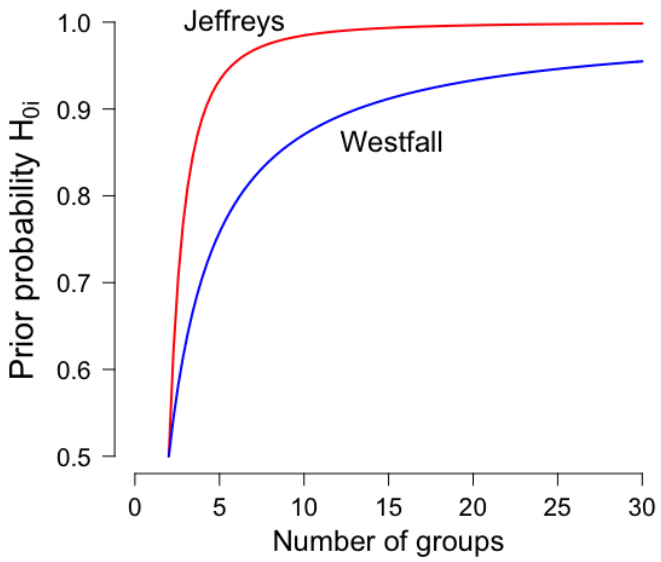

Figure 1. The prior model probability for an individual null hypothesis using the method of null control. This method assigns $\frac{1}{2}$ to the total prior probability of finding no effect in a set of pairwise comparisons. Jeffreys (red) is calculated assuming pairwise comparisons are independent, Westfall (blue) accounts for the dependency. Based on the number of groups the number of comparisons is obtained by the binomial coefficient, e.g., with 30 groups there are $\binom{30}{2} = 435$ pairwise comparisons.

these values do not have any direct link to the α threshold. It is unclear if an α of .05 coincides with anecdotal evidence, or decisive evidence. Consequently, instead of using these guidelines, our approach was to use the value for α and convert this to a new threshold for the Bayes factors. To this purpose we use the work of Sellke, Bayarri, and Berger (2001) who proposed a procedure to convert p-values to Bayes factors. Specifically, they provided a formula to obtain an upper bound on the Bayes factor BF_{A0} given p:

$$VS(p) = -\frac{1}{\mathrm{e}\, p \log(p)}, \tag{26}$$

when $p < \frac{1}{e}$. Now, given p must be smaller than .05 to be deemed significant, we may use this value to threshold the Bayes factor. Doing so gives us a value of $VS(.05) = 2.46$. Consequently, if BF_{A0} is larger than 2.46 we deem this as significant. We may now define false positives as the noise to noise comparisons for which we obtain $p < .05$ or $\text{BF}_{A0} > 2.46$.

Table 2
Commonly used interpretation categories of the Bayes factor (Lee & Wagenmakers, 2013, p. 105). An alternative table may be found in Kass and Raftery (1995). Shown is the Bayes factor for the marginal likelihood of the null model divided by that of an alternative model.

BF_{A0}	Interpretation
> 100	Extreme evidence for $\mathcal{H}_A$
$30 - 100$	Very strong evidence for $\mathcal{H}_A$
$10 - 30$	Strong evidence for $\mathcal{H}_A$
$3 - 10$	Moderate evidence for $\mathcal{H}_A$
$1 - 3$	Anecdotal evidence for $\mathcal{H}_A$
1	No evidence
$1/3 - 1$	Anecdotal evidence for $\mathcal{H}_0$
$1/10 - 1/3$	Moderate evidence for $\mathcal{H}_0$
$1/30 - 1/10$	Strong evidence for $\mathcal{H}_0$
$1/100 - 1/30$	Very strong evidence for $\mathcal{H}_0$
$< 1/100$	Extreme evidence for $\mathcal{H}_0$

The Bayes factor and p-value for each pairwise comparison were computed with the Bayesian and frequentist t-tests as implemented in the BayesFactor and stats packages. We recorded both the uncorrected outcome and the Bonferroni or null control corrected results. The combination of the t-test outcome with the α and BF_{A0} threshold allowed us to calculate the FDR. Recall that this requires dividing the number of false positive results by the total number of significant results ($\frac{V}{R}$ in Table 1). The False Omission Rate (FOR) was calculated by dividing the number of false negative results by the total number of non-significant results ($\frac{U}{m-R}$).

The data in our simulation was generated in R version 3.3.3 (R Core Team, 2017). We used a normal distribution with σ fixed to one or two. The normal distribution suits our purpose as sample normality is an assumption under the Student's t-test, as is equality of variances. We used four non-zero μ's and an increasing number of noise variables drawn from a zero-centered normal distribution. Noise was added in increments of 5 to a total of 30 noise variables. The parameters we chose for n and our signal μ's were based on the work of Bakker, van Dijk, and Wicherts (2012), they reported on 13 meta analytic studies performed in the field of psychology. The effect size across the studies ranged from .04 to 1.78. The parameters of our simulation reflected this range; we set the four μ's to .5, 1, 1.5 and 2. These values for μ — in addition to the zero-centered noise — allow the effect size to range from 0 to 2 and 0 to 1 for sd fixed to 1 and 2, respectively. The number of observations we choose were $n = 25$ and $n = 50$, these values approximate (1) the smallest

number of participants included in any of the studies and (2) the median number of included participants across studies. We intended to cover the studies with a low to average number of participants, a setting in which errors are more likely to occur.

We performed 500 repetitions for each unique combination of values for n and sd; the calculated FDR and FOR were averaged over the repetitions. The results of the simulation can be found in Figure 2. As expected, the FDR for the uncorrected frequentist t-test grows linearly with the number of added noise variables. A similar outcome is observed for the uncorrected Bayesian t-test, be it at a lower rate. The fact that the FDR is overall lower reflects that the Vovk-Sellke conversion from p-value to a Bayes factor provides an upper bound on the evidence. Ofttimes the Bayes factor will report less evidence for a difference between groups. After correction we find that the FDR rate diminishes for every combination of sample size and sd. This result is similar for both the Bayesian and frequentist approaches. The FDR is not quite zero, but approaches it closely as is evident in Figure 2 (lowest value reached is .10% for the Bayesian approach and .03% for the frequentist). The results show the effectiveness of these post-hoc methods as means of correcting for multiplicity. As is customary, when the Type-I error diminishes, the Type-II error increases; both procedures show a similar increase in the FOR after correction. The frequentist correction results in a slightly larger FOR across all situations.

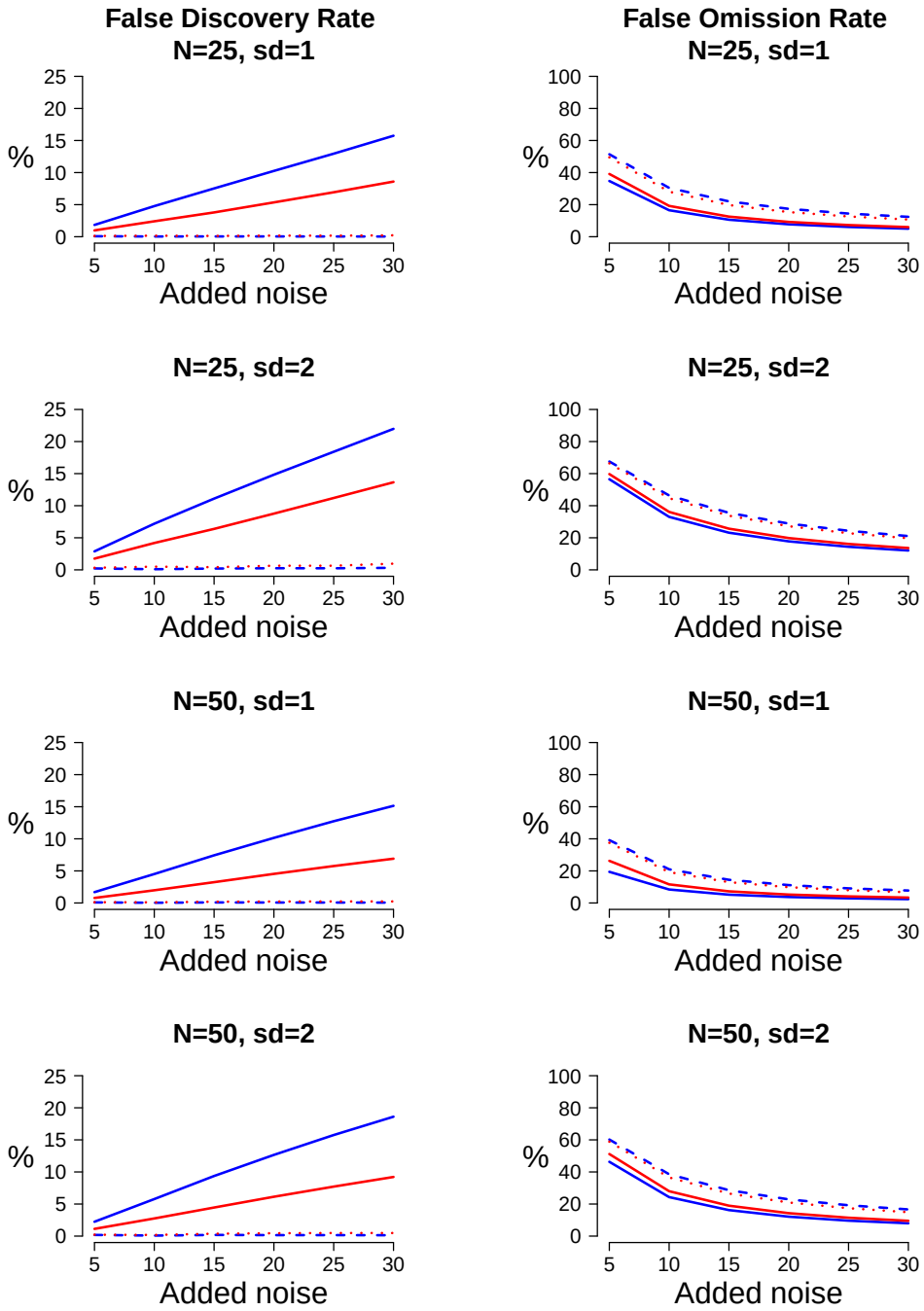

Figure 2. Pairwise comparisons of four signal variables ($\mu = .5$, 1, 1.5 and 2) with an increasing number of noise variables. Displayed are the FDR (left) and FOR (right) generated by frequentist and Bayesian t-tests. Each figure has uncorrected values: solid lines in blue (frequentist) and red (Bayesian), and corrected values: dashed blue (frequentist) and dotted red (Bayesian). FDR is calculated by dividing the number of false positives by the total number of positives. The FOR is calculated by dividing the number of false negatives by the total number of negatives. Significance threshold was $\alpha = .05$ and $\text{BF}_{A0} = -1/(\text{e} \cdot \alpha \cdot \log(\alpha)) \approx 2.46$. The displayed results were averaged over 500 iterations per condition.

Distribution of proportions

Before transitioning to pairwise comparisons we provide a quick introduction to the general method of SB (Scott & Berger, 2006, 2010). The method is geared towards regression and variable inclusion, and our aim is to adapt this method to pairwise comparisons.

Overview

In the case of variable inclusion or regression, each of m variables/coefficients can be independently included or excluded. This leads to a large possible model space with 2^m variations. To achieve multiplicity control it is important that the prior model probabilities are not uniformly assigned over models. Assigning 2^{-m} to each model is equivalent to assigning each variable/coefficient a prior probability of $\frac{1}{2}$ of being included. As we saw earlier in the example of variable inclusion, this provides no multiplicity control. Instead of assigning probability uniformly, Scott and Berger (2006, 2010) define a hierarchical structure over the model space. We will briefly discuss their implementation for regression.

Given a vector $\mathbf{y}$ of n responses and an n x m design matrix X, the regression model for the ith participant is given by

$$y_i = \beta_0 + X_{ij}\beta_j + ... + X_{im}\beta_m + \epsilon_i, \tag{27}$$

here $j = 1, ..., m$ and $i = 1, ..., n$. ϵ_i denotes a zero-centered noise term with unknown variance σ^2. All models include intercept term β_0. We denote the null model with only the intercept term as $\mathcal{M}_0$ and the full model with all covariates as $\mathcal{M}_m$. Each model is indexed by a binary vector γ of length m indicating the included and excluded regression coefficients:

$$\gamma_j = \begin{cases} 0, & \text{if } \beta_j = 0, \\ 1, & \text{if } \beta_j \neq 1. \end{cases} \tag{28}$$

The marginal likelihood used in their model was based on the null-based g-priors by Zellner (1986). As the regression analysis itself is not the focus of this study, we refer the interested reader to the appendix in Scott and Berger (2010) for further details.

We now turn to the part of their paper that is more relevant to us, the specification of prior model probabilities. Inclusion of every β_j is treated as a Bernoulli trial, where parameter q denotes the overall expected proportion of included coefficients:

$$p(\mathcal{M}_\gamma \mid q) = \prod_{j=1}^{m} q^{\gamma_j}(1-q)^{1-\gamma_j} \tag{29}$$

$$= q^{k_\gamma}(1-q)^{m-k_\gamma}, \tag{30}$$

where k_γ is the number of included coefficients in a model $\mathcal{M}_\gamma$ and m is the total number of coefficients under consideration. There are several options for specification of q, one option is to fix it to a specific value as we discussed earlier. But given that we have no subjective knowledge about q, we place a distribution on q instead; the usual choice is the Beta distribution:

$$q \sim \text{Beta}(a, b). \tag{31}$$

Although, in the Beta distribution itself it is again possible to express some belief in the proportion of included coefficients through the values of a and b. To this purpose Scott and Berger (2006) proposed the use of $\text{Beta}(1, b)$:

$$\pi(q) = b(1 - q)^{b-1}. \tag{32}$$

Figure 3 shows how this prior behaves for different values of b. Higher values would indicate that fewer coefficients are expected to be included. Setting b to 1 is the same as using a uniform distribution. For a further discussion of the influence of a and b as well as different priors on q see Li and Sivaganesan (2016).

Returning to the regular Beta distribution on q, we can obtain the prior model probabilities as follows:

$$p(\mathcal{M}_\gamma) = \int_0^1 p(\mathcal{M}_\gamma \mid q)\pi(q)\,\mathrm{d}q \tag{33}$$

$$= \frac{\text{B}(a + k_\gamma, b + m - k_\gamma)}{\text{B}(a, b)}, \tag{34}$$

where $\text{B}(\cdot)$ is the beta function. Now, if we have no prior knowledge and specify that $a = b = 1$, then the prior on q reduces to a uniform distribution. If we then integrate out q we find that this results in a simple partitioning of the prior probability over the dimensionality of the models:

$$p(\mathcal{M}_\gamma) = \frac{1}{m + 1}\binom{m}{k_\gamma}^{-1}. \tag{35}$$

The intuition behind this formula is that the prior model probability is first evenly shared across model classes of varying dimensionality (e.g., classes with models that have one β, two β's, etc.) and then within each of these classes across the number of possible configurations (e.g., only β_1 included, only β_2, etc.). Figure 4 shows how such partitions are made; Figure 5 shows the trade-off between the number of models and the probability each model receives.

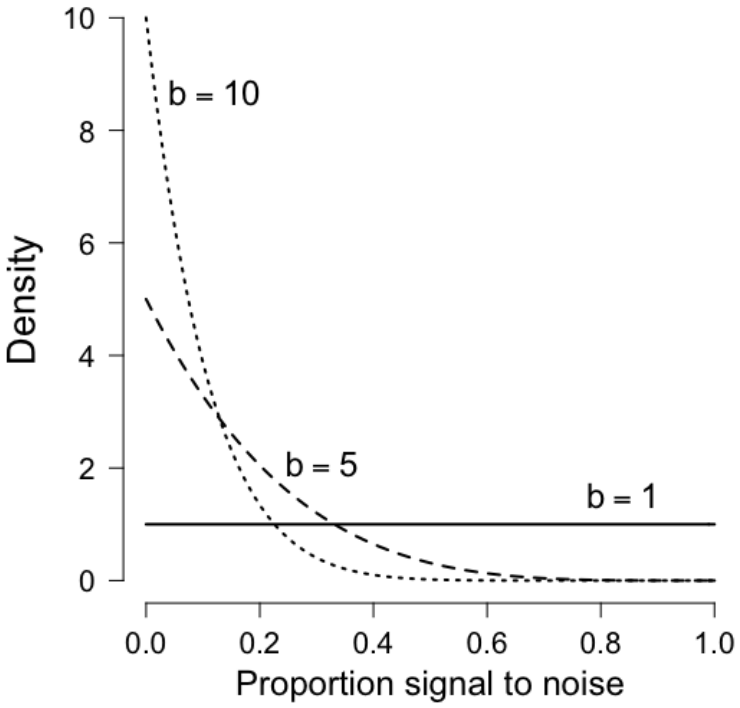

Figure 3. The expected proportion of signal to noise using a Beta$(1, b)$ distribution. A high value for b indicates an a priori expectation of mostly noise and few signal variables. If b is set to 1 the distribution defaults to a uniform distribution where every proportion value is equally likely.

Pairwise comparisons: Prior model probability

We now turn to pairwise comparisons in relation to the SB setup. Our interest shifts from inclusion of the regression coefficients β_i to the pairwise differences δ_i. Note also that Equation 35 changes slightly in this context, as there are not m pairwise comparisons we are interested in, but $k = \binom{m}{2}$. And so,

$$p(M_\gamma) = \frac{1}{k+1}\binom{k}{k_\gamma}^{-1}. \tag{36}$$

As an example, suppose we have 3 independent samples ($m = 3$) which leads to $k = \binom{3}{2} = 3$ pairwise comparisons. If we assume the δ's are also independent then the prior model probability would be partitioned as follows (where the subscript denotes the number

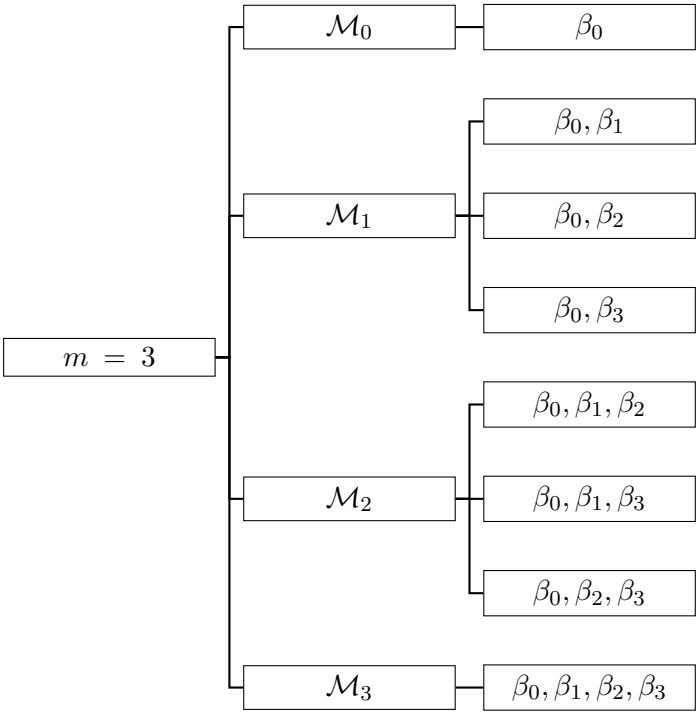

Figure 4. The model space for a regression model with three possible β coefficients. This diagram shows the intuition behind the partitioning of the prior model probability. From left to right: (1) number of regression coefficients under consideration, (2) classes of varying dimensionality and (3) possible model configurations within each class.

of included δ's):

$$p(\mathcal{M}_0) = \frac{1}{3+1}\binom{3}{0}^{-1} = \frac{1}{4}, \tag{37}$$

$$p(\mathcal{M}_1) = \frac{1}{3+1}\binom{3}{1}^{-1} = \frac{1}{12}, \tag{38}$$

$$p(\mathcal{M}_2) = \frac{1}{3+1}\binom{3}{2}^{-1} = \frac{1}{12}, \tag{39}$$

$$p(\mathcal{M}_3) = \frac{1}{3+1}\binom{3}{3}^{-1} = \frac{1}{4}. \tag{40}$$

There are three models in the dimension classes one and two, and so the total prior probability sums to one.

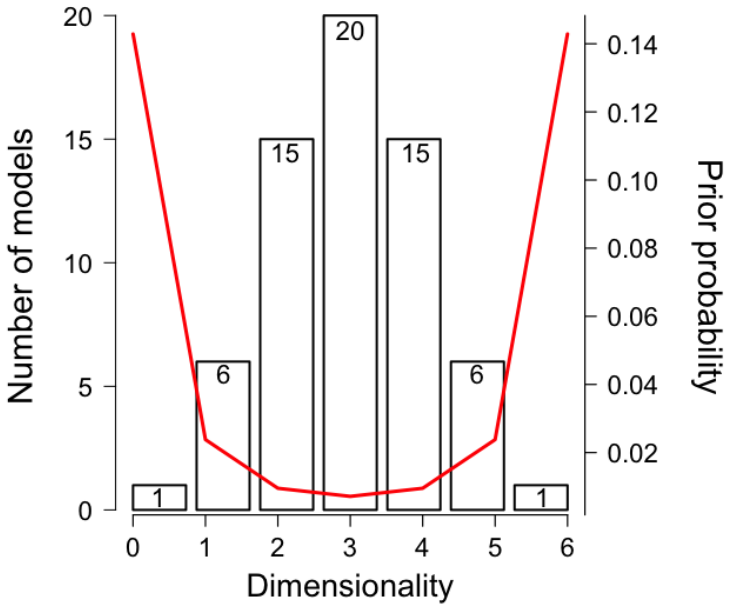

Figure 5. The trade-off between the number of models in a dimensionality class and the prior model probability each model receives. The dimensionality indicates the number of included β coefficients (out of a maximum of six) in a given model. The prior model probability is calculated using SB with a uniform prior on the proportion.

The problem with the method above is the same as for the Jeffreys correction: it does not take into account the dependency between pairwise comparisons. As such the partitioning of the prior probability over the model space is overly conservative. There is an additional concern that needs to be put to rest first; in the regression setup of SB, the assumption is that as the true number of non-zero regression coefficients remains constant in the face of ever increasing noise, q will tend to zero. In the case of pairwise differences, the true number of non-zero pairwise comparisons increases together with the added noise. What does this mean for q as more noise is added? To prove that this is of no concern, suppose we denote a fixed number of signal variables as x and an increasing number of noise variables y. Comparisons which show a true effect are those between two x variables and

between a x and y variable. The proportion of signal-to-noise q is then

$$q = \frac{xy + \binom{x}{2}}{\binom{x+y}{2}}. \tag{41}$$

Clearly, as the size of y increases while x remains fixed, Equation 41 tends to zero; consequently, the idea underlying regression holds for pairwise comparisons.

The problem of dependence is more difficult to address. To accomplish null control we could test the difference between any two groups in isolation of other groups; group A did not influence the comparison between B and C. With the SB method, however, we jointly model the differences between all groups. Just as we reasoned in terms of μ in the method for null control, we must again walk down this road — when we reason from the state of the groups (μ) we can infer which differences (δ) exist. One way to accomplish this is to treat groups and their differences as networks. In these networks nodes represent groups and edges represent differences between groups. To exemplify this procedure, we turn to the situation where $m = k = 3$. In Figure 6 we see that there are three possible dimension classes (zero, two or three edges). This stands in contrast with the four classes we get when independence is assumed, as we did in Equations 37-40. When we take into account the dependence between the δ's it becomes evident that we should only partition the prior model probability over possible models. Manually redistributing the previously calculated probability over three classes would result in:

$$p(\mathcal{M}_0) = \frac{1}{3}, \tag{42}$$

$$p(\mathcal{M}_2) = \frac{1}{9}, \tag{43}$$

$$p(\mathcal{M}_3) = \frac{1}{3}. \tag{44}$$

Where the class with dimensionality two still has the same three possible configurations, but $\mathcal{M}_1$ has been excluded. The total prior probability again sums to one.

It is evident that when we fail to consider dependency we assign probability to impossible models. This dependency can be shown graphically as a network. However, as the number of comparisons k grows, it quickly becomes infeasible to use these networks to determine model plausibility. A more convenient method is to determine logically constrained subsets of hypotheses based on equivalence relationships (Shaffer, 1986). This process is shown in the Table 3. It involves generating every possible equivalence relationship between the m groups. Say we reject the hypothesis $\mathcal{H}_{0_1}$ that some two groups are equal on the basis of an equivalence relationship. We then consider the largest collection of null hypotheses that could still be true, conditional on $\mathcal{H}_{0_1}$ being rejected. This process provides the possible

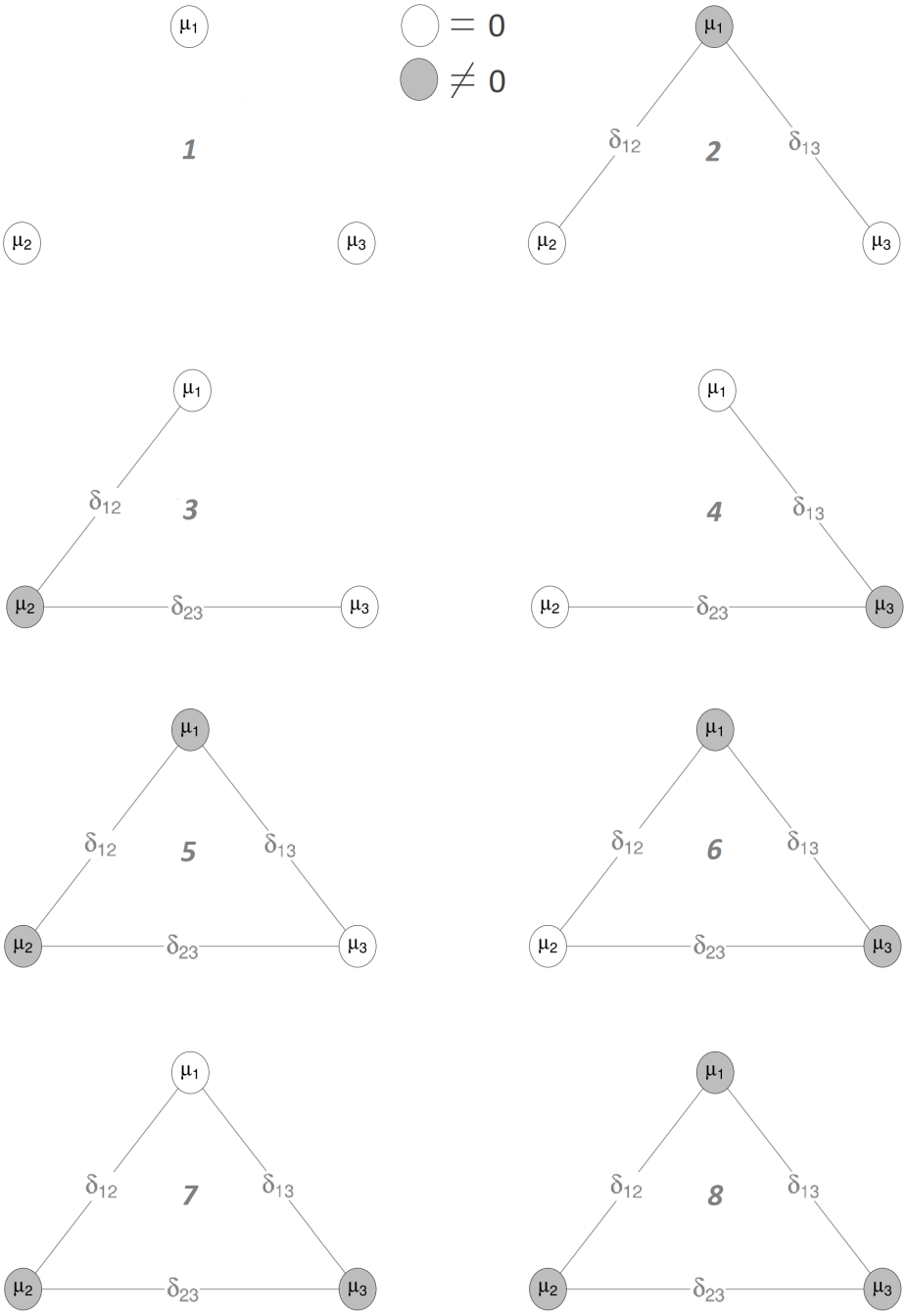

Figure 6. Pairwise comparisons for three sample means μ displayed as a network. Each μ is represented as a node and each difference between two means (δ) as an edge. The assumption here is that if (1) μ_1 and $\mu_2 \neq 0$ or (2) $\mu_1 \neq 0$ and $\mu_2 = 0$, then $\delta_{12} \neq 0$ and if μ_1 and $\mu_2 = 0$, then $\delta_{12} = 0$. The color indicates whether nodes are zero (white) or not (gray). We find eight distinct colorings of the nodes, but only five networks with unique edges (networks 1 to 5).

classes of hypotheses we may consider. Shaffer (1986) provided a recursion formula:

$$S(m) = \bigcup_{i}^{m} \left\{ \binom{i}{2} + x : x \in S(m-i) \right\}, \tag{45}$$

where $S(m)$ is the set of possible dimensionality classes for the null hypotheses, for $m \geq 2$ and given $S(0) = S(1) = \{0\}$. Note that we are not interested in the set of null hypotheses ($\delta = 0$), but the set of alternative hypotheses ($\delta \neq 0$). Consequently, we must subtract the set from k to obtain the classes for the alternative hypotheses (displayed in the third column of Table 3). To better understand the algorithm, we apply it to $m = 3$ groups. As the formula is recursive we first have to calculate $S(2)$:

$$S(2)_{j=1} = \left\{ \binom{1}{2} + S(1) \right\} = \{0\}, \tag{46}$$

$$S(2)_{j=2} = \left\{ \binom{2}{2} + S(0) \right\} = \{1\}, \tag{47}$$

$$S(2) = \{0, 1\}. \tag{48}$$

And so for $S(3)$ we get:

$$S(3)_{j=1} = \left\{ \binom{1}{2} + S(2) \right\} = \{0, 1\}, \tag{49}$$

Table 3
Finding possible subsets of the model space ($m = 4$) based on logically constrained relationships (Shaffer, 1986). For more details on the calculations in the rightmost column see Equation 59.

Partition	Number of $\delta = 0$	Number of $\delta \neq 0$	Number of configurations
$(\mu_1, \mu_2, \mu_3, \mu_4)$	$\binom{4}{2} = 6$	6 - 6 = 0	$\frac{4!}{4!} = 1$
$(\mu_1, \mu_2, \mu_3)(\mu_4)$	$\binom{3}{2} + \binom{1}{2} = 3$	6 - 3 = 3	$\frac{4!}{3!1!} = 4$
$(\mu_1, \mu_2)(\mu_3, \mu_4)$	$\binom{2}{2} + \binom{2}{2} = 2$	6 - 2 = 4	$\frac{4!}{(2!)^3} = 3$
$(\mu_1, \mu_2)(\mu_3)(\mu_4)$	$\binom{2}{2} + \binom{1}{2} + \binom{1}{2} = 1$	6 - 1 = 5	$\frac{4!}{(2!)^2(1!)^2} = 6$
$(\mu_1)(\mu_2)(\mu_3)(\mu_4)$	$\binom{1}{2} + \binom{1}{2} + \binom{1}{2} + \binom{1}{2} = 0$	6 - 0 = 6	$\frac{4!}{(1!)^4 4!} = 1$

$$S(3)_{j=2} = \left\{ \binom{2}{2} + S(1) \right\} = \{1\}, \tag{50}$$

$$S(3)_{j=3} = \left\{ \binom{3}{2} + S(0) \right\} = \{3\}, \tag{51}$$

$$S(3) = \{0, 1, 3\}. \tag{52}$$

We then proceed to subtract this set from k — which for $m = 3$ is $\binom{3}{2} = 3$:

$$k - S(3) = 3 - \{0, 1, 3\} = \{0, 2, 3\}. \tag{53}$$

And so we obtain the possible classes (zero, two and three possible δ's) as shown in Figure 6. With the algorithm we have a way of obtaining the set of classes that are logically possible.

Previously, for the method of null control, we showed the difference between ignoring and correcting for dependence. Similarly, it would be interesting to know how many models are pruned when we take the dependency into account with SB. For this we need the number of distinct hypotheses — and so, how many models — a given set of classes provides. For example, looking at the difference between Equations 37-40 and 42-44 we find a reduction of three models across the classes. As it turns out, the total number of distinct hypotheses we obtain for some m is equal to the Bell number (Berry & Christensen, 1979). More information about the Bell number can be found in Box 1. Its recursive formula is

$$Bn_{m+1} = \sum_{i=0}^{m} \binom{m}{i} Bn_i. \tag{54}$$

So for $m = 3$ we find $Bn(3) = 5$, which is equal to the unique configurations of edges — although there are eight networks, the last four have identical edges — we saw in Figure 6. With Shaffer's formula and the Bell number we can compute the constrained set of possible models and classes; the results are displayed in Figure 7. We observe a strong reduction in size when considering the constrained subset instead of the full set, which shows the importance of not assuming independence between hypotheses.

It would seem that the only remaining challenge is to integrate Shaffer's subsets with SB. This is fairly straightforward, as we are interested in a simple redistribution of the prior model probabilities over the new model space. This redistribution is proportional to the size difference between the logically constrained subset $S(m)$ and the unconstrained model space. When we include this proportion term in SB we get:

$$p(\mathcal{M}_\gamma) = \frac{k+1}{|S(m)|} \cdot \frac{1}{k+1} \binom{k}{k_\gamma}^{-1} \tag{55}$$

$$= \frac{1}{|S(m)|}\binom{k}{k_\gamma}^{-1}, \tag{56}$$

where $k_\gamma \in k - S(m)$. However, the problem of dependence is not solved entirely by taking into account the restrained subset of classes. To exemplify this, consider $m = 4$ and so $k = \binom{4}{2} = 6$. From Shaffer we obtain the set $S(4) = \{0, 3, 4, 5, 6\}$ — see also Table 3. The total number of models in this set is $Bn(4) = 15$. Now, if we compute the prior model probability for a model that has four out of the six possible δ's:

$$p(\mathcal{M}_4) = \frac{1}{5}\binom{6}{4}^{-1} = \frac{1}{75}, \tag{57}$$

we find that every model in this class receives $\xi_A = \frac{1}{75}$. Recall that the SB method first distributes prior probability over the possible classes and then over the different model configurations within classes. With 5 classes, it would mean that $\mathcal{M}_4$ has 15 different configurations. However, the Bell number showed us there were a total of 15 models across all sets; this would mean that for $k = 6$ we only have models with 4 δ's. Obviously this is

Box 1. The idea behind the Bell number.

The Bell number, named after Eric Temple Bell, counts the ways a set of elements can be split into subsets (Bell, 1934). The splits of a set must result in nonempty, mutually disjoint subsets. The union of the subsets is the set once more. The first 10 Bell numbers are 1, 1, 2, 5, 15, 52, 203, 877, 4140 and 21147. We exemplify the basic mechanism with a set containing the three elements A, B and C. The different ways we can split this set into subsets are the following:

(1) { {A}, {B}, {C} },
(2) { {A}, {B, C} },
(3) { {B}, {A, C} },
(4) { {C}, {A, B} },
(5) { {A, B, C} }.

Consequently, based on a set of size three, we find five different ways of splitting the set.

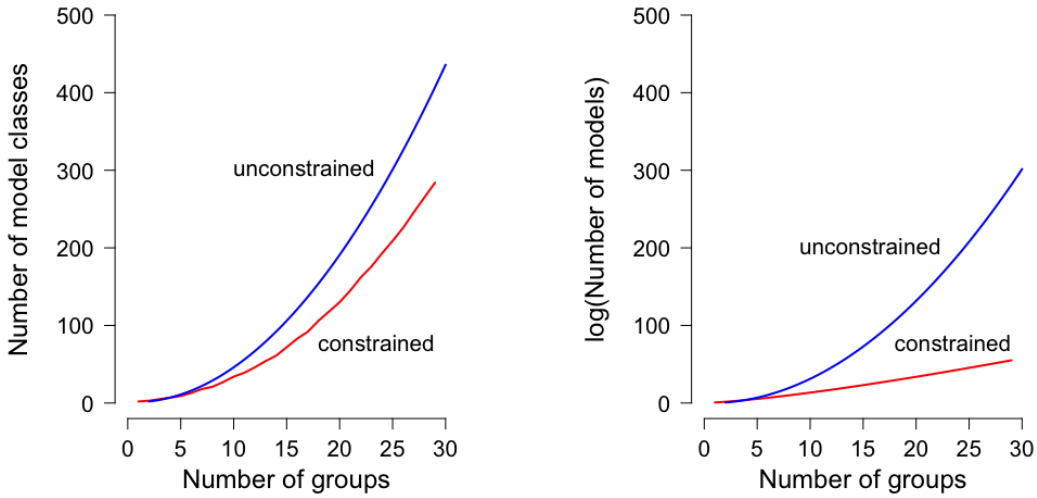

Figure 7. The number of model classes (left) and the size of the model space (right) as a function of the number of groups under consideration. The unconstrained number of classes and models (blue) assumes independence between pairwise comparisons. When we account for dependence between comparisons we obtain the constrained sets (red). These constrained sets are calculated by using equivalence relationships of the groups underlying the comparisons (equal to the Bell number). The figure shows an increasing difference between the two lines, meaning an increase in logically impossible models and classes.

not the case, and indeed, if we look at $\mathcal{M}_5$:

$$p(\mathcal{M}_5) = \frac{1}{5}\binom{6}{5}^{-1} = \frac{1}{30}, \tag{58}$$

it becomes clear there are an additional $\binom{6}{5} = 6$ models. The 15 models in $\mathcal{M}_4$ (and 6 models in $\mathcal{M}_3$) only exist when we once more assume independence between each δ. Given independence there are $\binom{6}{4} = 15$ ways of choosing 4 δ's out of 6. It is clear that Equation 56 does not solve the problem of dependence.

It is only possible to compute the number of model configurations in a class when we know the underlying state of the μ's. In Table 3 we find that the state of the μ's underlying class $\mathcal{M}_4$ is based on $(\mu_1 = \mu_2) \neq (\mu_3 = \mu_4)$. Now, using the logic of partitioning a set of size m in u unordered subsets with r elements (where $u \cdot r = m$), we find that there are three possible configurations:

$$\frac{m!}{u!(r!)^u} = \frac{4!}{2!(2!)^2} = 3. \tag{59}$$

The partitions in this specific case are: $(\mu_1 = \mu_2) \neq (\mu_3 = \mu_4)$, $(\mu_1 = \mu_3) \neq (\mu_2 = \mu_4)$ and $(\mu_1 = \mu_4) \neq (\mu_2 = \mu_3)$. Each of these configurations leads to four non-zero δ's. In the final column of Table 3 we can find the number of configurations for the remaining classes. When we sum over this column we find the 15 models we obtained earlier with the Bell number. It is clear that the issue of dependence is not yet overcome by simply combining SB with Shaffer. Only for $m = k = 3$ does Equation 56 provide a satisfactory answer, for $m > 3$ additional computations are required. Fortunately, it seems these computations can be made by a combination of Shaffer's method and the additional combinatorics shown in Equation 59:

$$p(\mathcal{M}_\gamma) = \frac{1}{|S(m)|} \left(\frac{m!}{u!(r!)^u} \right)^{-1}. \tag{60}$$

Note that Equation 60 requires us to know the u subsets and the r elements in the subsets that contribute to a class $\mathcal{M}_\gamma$. If we have this information, we know how to assign prior model probabilities while accounting for dependency.

Pairwise comparisons: Marginal likelihood

Giving the prior model probabilities a rest we turn to the marginal likelihood. Earlier we defined the t-test and were able to utilize this for the method of null control. However, this is no longer possible, as we will often have to compare more than two groups simultaneously. This becomes apparent when we look at $m = k = 3$; we will not be able to calculate the evidence for the model in class $\mathcal{M}_3$ with a t-test, as this involves $\mu_1 \neq \mu_2 \neq \mu_3$. Now, we may utilize the model as specified by Scott and Berger (2006), however, it uses a different set of priors than the t-test. To keep our implementation consistent we turn to the Bayesian one-way ANOVA instead (Rouder, Morey, Speckman, & Province, 2012). It is a direct extension to the t-test in the sense that an ANOVA on two groups returns identical results to the t-test. The linear model is

$$y_{ij} = \mu + \omega_i + \epsilon_{ij}, \tag{61}$$

for $i = 1, ..., m$ levels in a factor, with $j = 1, ..., n$ observations in each. μ is the grand mean and ω_i the effect of the ith level of the factor; ϵ_i is the zero-centered error term. A problem with this model is that more parameters contribute to the mean of each level, than there are actual levels. The result of this problem is that parameters cannot be uniquely identified. The approach Rouder et al. (2012) took was to add a sums-to-zero constraint ($\sum \omega_i = 0$).

Just as for the t-test the model terms ω_i are reparameterized in terms of the effect size $d_i = \frac{\omega_i}{\sigma}$. Priors for parameters common in all models receive the same uninformative

prior:

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{62}$$

Models that include a d parameter are specified much like the t-test:

$$d_i \sim \text{Normal}(0, \sigma_d^2), \tag{63}$$

$$\sigma_d^2 \sim \text{inverse chi-square}(1). \tag{64}$$

Which, as noted earlier, reduces to a Cauchy distribution after integrating out the variance

$$d_i \sim \text{Cauchy}. \tag{65}$$

Of course, we are interested in inferences on δ and not on d. Using the ANOVA model as is, would provide us no information about any particular pairwise difference. The ANOVA only tests whether all groups are equal or not. But then, how should the linear model be defined to allow for inferences on δ and not d? Furthermore, should the δ's then sum to zero? Would a sums-to-zero constraint have any real interpretation in this context? At present we have no answers to these questions. Fortunately, a different route exists which avoids these pitfalls — relabeling.

Let's again consider the situation where $m = k = 3$. Now, say that our hypotheses were the following: $\delta_{12} = 0$ and both $\delta_{13}, \delta_{23} \neq 0$. In essence we would be saying that because $\delta_{12} = 0$ there exists no difference between group 1 and group 2. So, referring to group 1 as group 2 and vice versa does not change any of the inferences in this set of hypotheses. If the two are interchangeable in this regard, assigning all observations of these two groups to one joint group would also be valid. The only comparison of interest would then be $\delta_{\{12\}3}$, which can be evaluated with a one-way ANOVA (or simply a t-test in this case). This is in line with the notion of Rouder et al. (2012, p. 363):

> "In ANOVA designs, researchers are sometimes concerned about additional contrasts, such as whether any two levels differ. For instance suppose a factor has three levels and the main-effect Bayes factor indicates that the full model is preferred to the null model. Then, three intermediate models may be proposed where any two levels equal each other. Each of these models can be implemented with a simple two-column design matrix and tested with the above methodology. The resulting pattern of Bayes factors across these models, as well as that across the full model, may be compared in analysis."

This method, where equality between means is obtained by a simple relabeling, has been applied in a number of Bayesian studies that looked at pairwise comparisons (e.g., Gopalan

& Berry, 1998; Neath & Cavanaugh, 2006). If we return to our $m = k = 3$ scenario, its Bayes factors can then be calculated as follows:

$$\mathrm{BF}_{00}^{\mathcal{M}_0} = 1, \tag{66}$$

$$\mathrm{BF}_{A0}^{\mathcal{M}_2} = \begin{cases} \dfrac{p(\mathbf{y} \mid \mu_{\{12\}}, \mu_3)}{p(\mathbf{y} \mid \mu_1 = \mu_2 = \mu_3)}, & (67) \\ \dfrac{p(\mathbf{y} \mid \mu_{\{13\}}, \mu_2)}{p(\mathbf{y} \mid \mu_1 = \mu_2 = \mu_3)}, & (68) \\ \dfrac{p(\mathbf{y} \mid \mu_{\{23\}}, \mu_1)}{p(\mathbf{y} \mid \mu_1 = \mu_2 = \mu_3)}, & (69) \end{cases}$$

$$\mathrm{BF}_{A0}^{\mathcal{M}_3} = \frac{p(\mathbf{y} \mid \mu_1, \mu_2, \mu_3)}{p(\mathbf{y} \mid \mu_1 = \mu_2 = \mu_3)}. \tag{70}$$

Pairwise comparisons: An example

We work out a simple example to demonstrate the SB method in combination with the Rouder et al. (2012) framework — we use a dataset with $m = 4$ groups and $k = \binom{4}{2} = 6$ pairwise comparisons between groups. Note that this situation coincides with the situation we showed in Table 3 and for some calculations we refer to this table. The 4 groups in the dataset each had $n = 50$ observations and were drawn from a normal distribution with a standard deviation of 1. Only the first group had a non-zero μ, which was set to .5.

We first turn to the prior probabilities. With Shaffer we get the set of possible model classes $S(4) = \{0, 3, 4, 5, 6\}$. Using Equation 59 we find that models are distributed across the classes as follows (for calculations see the last column in Table 3): one in $\mathcal{M}_0$, four in $\mathcal{M}_3$, three in $\mathcal{M}_4$, six in $\mathcal{M}_5$ and one in $\mathcal{M}_6$. Consequently, the SB corrected prior model probabilities for each class are obtained with Equation 60:

$$p(\mathcal{M}_0) = \frac{1}{5}\frac{4!}{4!} = \frac{1}{5}, \tag{71}$$

$$p(\mathcal{M}_3) = \frac{1}{5}\frac{4!}{3!1!} = \frac{1}{20}, \tag{72}$$

$$p(\mathcal{M}_4) = \frac{1}{5}\frac{4!}{(2!)^3} = \frac{1}{15}, \tag{73}$$

$$p(\mathcal{M}_5) = \frac{1}{5}\frac{4!}{(2!)^2(1!)^2} = \frac{1}{30}, \tag{74}$$

$$p(\mathcal{M}_6) = \frac{1}{5}\frac{4!}{(1!)^4 4!} = \frac{1}{5}. \tag{75}$$

The uncorrected prior probabilities for each model are obtained by simply dividing the

probability equally among the models. The total number of possible models can be found with the Bell number, $Bn(4) = 15$. And so each model regardless of its class receives $\frac{1}{15}$ in the uncorrected situation.

The Bayes factor BF_{A0} for each model was obtained with a Bayesian ANOVA from the BayesFactor package (Morey & Rouder, 2015). Each of these models was defined by relabeling groups according to the left column in Table 3. As an example, $(\mu_1 = \mu_2) \neq (\mu_3 = \mu_4)$ resulted in the relabeled groups $\{12\}$ and $\{34\}$ with the Bayes factor

$$\text{BF}_{A0}^{\mathcal{M}_{4,z}} = \frac{p(\mathbf{y} \mid \mu_{\{12\}}, \mu_{\{23\}})}{p(\mathbf{y} \mid \mu_1 = \mu_2 = \mu_3 = \mu_4)}, \tag{76}$$

where z signifies this particular model configuration in class $\mathcal{M}_4$. The relabeling method provided us with 15 Bayes factors. To obtain the posterior odds, the prior model probabilities calculated earlier were turned to prior odds and then multiplied by the Bayes factors. Continuing with the example of model z in class $\mathcal{M}_4$:

$$\frac{p(\mathcal{M}_{4,z} \mid \mathbf{y})}{p(\mathcal{M}_0 \mid \mathbf{y})} = \frac{p(\mathcal{M}_4)}{p(\mathcal{M}_0)} \text{BF}_{A0}^{\mathcal{M}_{4,z}}, \tag{77}$$

where the subscript denoting the specific model in $\mathcal{M}_0$ was omitted, as there is only one possibility. Subsequently, we computed the posterior probability for each model; to do so we divided the posterior odds of each model by the sum of all posterior odds. For model z we compute

$$p(\mathcal{M}_{4,z} \mid \mathbf{y}) = \frac{\frac{p(\mathcal{M}_{4,z}|\mathbf{y})}{p(\mathcal{M}_0|\mathbf{y})}}{\sum_{i=0}^{|S(4)|} \sum_{j=1}^{|\mathcal{M}_i|} \frac{p(\mathcal{M}_{i,j}|\mathbf{y})}{p(\mathcal{M}_0|\mathbf{y})}}, \tag{78}$$

where $|S(\cdot)|$ is the number of model classes obtained with Shaffer and $|\mathcal{M}_i|$ is the number of models in class i. Note that this conversion from Bayes factors to posterior probabilities may be employed when all Bayes factors have the same denominator model and so each Bayes factor states the evidence relative to that fixed model.

The final step required us to obtain posterior inclusion probabilities for the effects. This required a sum of the posterior probabilities of the models that included a given effect. As an example, model z contributed to the evidence for the effects $\delta_{13}, \delta_{14}, \delta_{23}$ and δ_{24}. In a similar way we obtained the prior inclusion probabilities by summing over the prior model probabilities of relevant models. Calculating the change from prior inclusion to posterior inclusion probabilities gave us the inclusion Bayes factors.

The entire procedure was repeated 500 times to reduce the influence of the random data sampling on the results. The prior inclusion probabilities, median posterior inclusion probabilities and inclusion Bayes factors can be found in Table 4. To determine which effects should be included in the model a common cut-off value for the posterior inclusion proba-

Table 4
Inclusion probabilities for the differences between four groups. The first group was normally distributed with $\mu = .5$, while the other groups were normally distributed around zero. Uncorrected models received equal probability, corrected models received probability according to the SB method. We used the median outcome over 500 repetitions for the posterior inclusion probability and inclusion Bayes factor.

	Uncorrected			Corrected		
Effect	Prior incl.	Post. incl.	Inclusion BF	Prior incl.	Post. incl.	Inclusion BF
δ_{12}	.667	.910	1.366	.600	.853	1.421
δ_{13}	.667	.910	1.365	.600	.854	1.424
δ_{14}	.667	.911	1.366	.600	.856	1.427
δ_{23}	.667	.417	.626	.600	.412	.686
δ_{24}	.667	.424	.636	.600	.409	.682
δ_{34}	.667	.417	.626	.600	.413	.689

bilities is .5 Using this cut-off, it is clear that the relabeling procedure correctly identified the true δ's: all comparisons to the first group have posterior inclusion probabilities greater than .5 — both uncorrected and corrected. The noise to noise comparisons failed to reach .5 and were correctly rejected. When we then compare the SB method and the uncorrected situation, we note a couple of differences. Firstly, the prior inclusion probabilities are lower, meaning that each δ receives less probability of being included a priori. Secondly, we find a difference between the posterior inclusion probabilities. There is a decrease in the posterior probabilities for all δ's under SB. Clearly, this is what would be expected when we correct for multiplicity. However, no conclusions may be drawn from this; more noise groups would have to be added to see if the current implementation of the SB method truly corrects for multiplicity.

Discussion

For this thesis we set a number of goals; first and foremost we wanted to make it easier to deal with the multiplicity problem in Bayesian statistics — specifically in relation to pairwise comparisons after a one-way ANOVA. We evaluated two methods both based on adjusting prior model probabilities. The first method, null control, was shown to be an effective way of dealing with multiplicity. Extending Jeffreys' method by accounting for dependency made the method less conservative, while retaining its underlying idea of keeping the probability of no effects to .5. It proved as effective as a frequentist Bonferroni correction, while being slightly less conservative. For the second method we examined, a successful translation to pairwise comparisons proved more complicated. We showed the importance of dealing with dependence in this context, as a large number of models are otherwise considered which are logically impossible. We ultimately succeeded in adapting

the method by different means of combinatorics. However, while the method of null control has been implemented in JASP and can be used for an arbitrary number of levels in a factor, this is not the case for the SB method. If we are working with m levels, we must first obtain all possible equivalence relationships. Only then, based on these relationships and the partitions they create, can we obtain the prior model probabilities for each model. Consequently, it will require additional work to make the SB method more generally applicable. And, more importantly, while the validity of SB has been shown in various different applications (Bogdan et al., 2008; Carvalho & Scott, 2009; Q. Li & Shang, 2015; Mitra et al., 2017; Scott & Berger, 2006, 2010), we were unable to do so as of yet in the ANOVA context. As a consequence, we could also not numerically compare both methods in this study.

One difference between the two methods is rather apparent, though, without extensive simulation; namely, the assumption we make when we account for dependence. To achieve null control under dependence, we assumed each μ_i may either be equal to the grand mean or be different altogether: a value drawn from a continuous distribution. As such if two groups are not equal to the grand mean they may not be equal to each other. This assumption is more stringent than the one we considered for SB. In obtaining our subset of models through equivalence relationships, we do allow two groups to equal each other, even when they are different from the grand mean. Which approach is better is debatable, however, it would be interesting to explore the impact of this difference between the methods.

There are two notions about the methods that we feel are necessary to put forth at this point. The first is that we used the idea of SB, but not necessarily the common implementation. SB as proposed by Scott and Berger (2006, 2010) can accommodate any prior on the proportion of included effects (Li & Sivaganesan, 2016). If this prior is uniform their implementation and ours coincide, prior probability is distributed over the classes and then over the configurations in the classes. However, as our implementation deviates because of the dependency problem, we cannot easily make the transition to a different prior on the proportion of included effects. This may not be a large obstacle in the current context, as post-hoc comparisons following a one-way ANOVA are often exploratory, without specific hypotheses defined a priori. The second notion is more philosophical in nature and relates to the method of null control. The core idea behind it is that we value — and therefore must protect — the possibility of an invariance among our set of hypotheses. However, if we are to believe Cohen (1994) then true invariance does not exist, nothing is ever exactly equal to zero. This is a view shared by Schmidt (1992) who, as a result, claims Type-I error cannot occur and simply hinders the scientific focus on development of cumulative knowledge. Should we then assign this event of complete invariance such high probability? Maybe not, but as we saw in our simulation, failing to account for the possibility of invariance

leads to an increased number of false positives. This is an issue that is of great concern, considering the growing consensus that reliability of published research is poor at best. In a recent survey by Nature involving 1,576 researchers a total of 52% respondents reported that science is suffering a significant reproducibility crisis (Baker, 2016). It is clear that the number of published false positives must be addressed for researchers to regain confidence in the scientific principles. Methods such as null control make a necessary contribution to this cause.

In summary, we looked at the implementation of two methods for multiplicity control. We showed that the method of null control is able to protect against Type-I errors and is on a similar footing to the famous frequentist Bonferroni procedure. The SB method requires some work, but the essentials have been given in this thesis. Null control will be made available through the graphical software JASP. It is another tool researchers can use to to guard themselves against false positive results. With some luck this will remove the necessity to do any additional research on these poor, dead salmon.

References

Abramovich, F., & Angelini, C. (2006). Bayesian maximum a posteriori multiple testing procedure. *Sankhyā: The Indian Journal of Statistics*, *68*, 436–460.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bell, E. T. (1934). Exponential numbers. *The American Mathematical Monthly*, *41*, 411–419.

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 (Statistical Methodology)*, *57*, 289–300.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009, June). *Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction.* Poster presented at the Human Brain Mapping Conference, San Fransisco, CA.

Berry, D. A., & Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of dirichlet processes. *The Annals of Statistics*, *7*, 558–568.

Bogdan, M., Ghosh, J. K., & Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *Institute of Mathematical Statistics Collections*, *1*, 211–230.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.

Carvalho, C. M., & Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, *96*, 497–512.

Chen, J., & Sarkar, S. K. (2004). Multiple testing of response rates with a control: A Bayesian stepwise approach. *Journal of Statistical Planning and Inference*, *125*, 3–16.

Clyde, M. (2017). BAS: Bayesian Adaptive Sampling for Bayesian model averaging (R package version 1.4.6) [Computer software]. Retrieved from `https://CRAN.R-project.org/package=BAS`

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, *38*, 143–151.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript. Retrieved from `http://www.stat.columbia.edu/ gelman/research/unpublished/p_hacking.pdf`

Gopalan, R., & Berry, D. A. (1998). Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association*, *93*, 1130–1139.

Hochberg, Y., & Tamhane, A. (1987). *Multiple comparison procedures.* New York, NY: Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

JASP Team. (2017). JASP (Version 0.8.1.2)[Computer software]. Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *165*, 161–198.

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-016-1221-4

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.

Li, & Sivaganesan, S. (2016). On the role of the prior in multiplicity adjustment. *Journal of Statistical Theory and Practice*, *10*, 263–290.

Li, Q., & Shang, J. (2015). A Bayesian hierarchical model for multiple comparisons in mixed models. *Communications in Statistics - Theory and Methods*, *44*, 5071–5090.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, *5*, 269–287.

Mitra, R., Mueller, P., & Ji, Y. (2017). Bayesian multiplicity control for multiple graphs. *The Canadian Journal of Statistics*, *45*, 44–61.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.12-2) [Computer software]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor`

Neath, A. A., & Cavanaugh, J. E. (2006). A Bayesian approach to the multiple comparisons problem. *Journal of Data Science*, *4*, 131–146.

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.3.2) [Computer software]. R Foundation for Statical Computing, Vienna, Austria. Retrieved from `https://www.R-project.org/`

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*, 1237–1282.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*, 2144–2162.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587–2619.

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826–831.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *10*, 681–690.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*, 9440–9445.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*, 99–114.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, *92*, 299–306.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*, 419–427.

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgment and Decision Making*, *11*, 174–184.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior

distributions. In P.K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233-243).