Structured testing of 2×2 factorial effects: an analytic plan requiring fewer observations

Dylan S. Small[*], Kevin G. Volpp, Paul R. Rosenbaum
Department of Statistics and Department of Medicine
University of Pennsylvania, Philadelphia, PA 19104
18 January 2011

Abstract

In a randomized 2×2 factorial trial, more than one hypothesis is to be tested, so some method must be used to control the probability of falsely rejecting at least one true hypothesis. We contrast familiar elementary methods of controlling the family-wise error rate based on the Bonferroni-Holm procedure with a less familiar but equally elementary form of structured testing associated with the large class of procedures that descend from the closed testing approach of Marcus, Peritz and Gabriel. In a range of plausible situations, giving priority to main effects in structured testing typically yielded greater power to detect main effects for a given sample size or reduced sample size for a given power; it also permited testing for interaction when main effects are found.

1 Introduction: smaller required sample sizes through better analytic plans

The randomized 2×2 factorial design is widely used in clinical trials, permitting two treatments to be studied efficiently in a single trial. For two examples, see Brown et al. (2001) and Stone et al. (2002), and for general discussion, see Byar and Piantandosi (1985), Stampfer et al. (1985), and Ellenberg, Finkelstein and Schoenfeld (1982). In such trials, the main effects of the two treatments are often of primary concern, and the 2×2

[*] Dylan S. Small is associate professor and Paul R. Rosenbaum is professor in the Department of Statistics of the Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 US, dsmall@wharton.upenn.edu. Kevin G. Volpp is associate professor at the University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, and director of the Center for Health Incentives at the Leonard Davis Institute of Health Economics. This work was supported by grants SES-0849370 and SES-0961971 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation.

factorial design is used to determine whether either or both treatments are effective; in addition, some information is provided about interactions between the treatments. To interpret such a trial, more than one hypothesis is tested, and some method must be used to control the probability that a true null hypothesis is falsely rejected. How should this probability of false rejection be controlled?

The probability of falsely rejecting at least one true null hypothesis, say π, is the family-wise error rate. One familiar way to obtain $\pi \leq \alpha$ when performing K tests is to apply the Bonferroni inequality, rejecting the k^{th} of K hypotheses if its P-value is less than or equal to α/K, and Holm's (1979) method is similar but offers an improvement in power. Another way to obtain $\pi \leq \alpha$ is to use one of the many methods of structured testing that descend from the paper on closed testing by Marcus, Peritz and Gabriel (1976) in which testing is given a structure in which hypotheses are tested each at level α and testing terminates for certain patterns of acceptances. For a few of the many such structured descendents of closed testing, see Shaffer (1986), Koch and Gansky (1996), Bauer and Kieser (1996), Bauer (1997), Hsu and Berger (1999), Westfall and Krishen (2001), Finner and Strassburger (2002), Hommel and Kropf (2005), Hommel, Bretz and Mauer (2007), Meinshausen (2008), Rosenbaum (2008), Ehm, Kornmeier and Heinrich (2010), Finos and Farcomeni (2010) and Goeman and Solari (2010). A common feature of these descendents is that they permit somewhat more flexibility in the structure of the testing than was possible in the original version of closed testing. For instance, the methods of Bauer and Kieser (1996), Hsu and Berger (1999) and Rosenbaum (2008) may test infinitely many hypotheses.

One of us recently submitted a proposal to NIH for a 2×2 factorial randomized clinical trial to evaluate two treatments that provide incentives for cognitive exercise in the elderly with a view to diminishing or delaying dementia; see Willis et al. (2006) and Papp et al. (2009) for discussion of related studies. In the process of preparing that proposal, we performed a few power or sample size calculations, and in particular we compared some fairly conventional uses of the Bonferroni-Holm procedure with a simple version of structured testing tailored to the 2×2 factorial design. An abbreviated and simplified version of this comparison appears in Table 1. Although there is no uniformly most powerful procedure in this context, the power comparison generally favored the structured testing approach, yielding higher power for the same sample size or lower sample size for the same power. This is in contrast to typical textbooks, typical courses in statistical methods, and typical scientific practice in which procedures such as the Bonferroni-Holm method are prominent and structured testing is infrequent. Our interpretation of Table 1 is that a change in emphasis towards structured testing may be appropriate. We focus on the 2×2 factorial design because it was the original motivation for this work and because it is sufficiently simple that we may present a fairly exhaustive power comparison.

Outside of randomized experiments, in nonrandomized or observational studies, structured testing has certain additional advantages when studying the sensitivity of conclusions to departures from random assignment. For discussion, see Rosenbaum and Silber (2009) and Rosenbaum (2010, §19).

2 Methods: three analytic plans that control the probability of falsely rejecting a true hypothesis

We contrast the power of three analytic plans to reach various conclusions in the 2×2 factorial. All three plans control the chance of falsely rejecting any true hypothesis, but they do this in different ways. We describe the plans in terms of level α, $0 < \alpha < 1$; in practice, this is commonly $\alpha = 0.05$. The first two are entirely standard plans, using Holm's (1979) improvement of the familiar Bonferroni inequality; see Hochberg and Tamhane (1987, §4) or Lehmann and Romano (2005, §9). The two standard plans differ in that plan I tests for main effects and interaction, allowing for three tests, while plan II tests only for main effects, allowing for two tests, so plan II has more power to detect main effects and no possibility of detecting an interaction. The third procedure (III) uses one simple version of structured testing. Each of the three analytic plans controls the probability of false rejection — though several hypotheses are tested, the chance of falsely rejecting at least one true hypothesis is at most α — however, they do this in different ways, and in particular, they test slightly different hypotheses. In describing hypotheses, logical notation is used, so $H \wedge H'$ is the hypothesis that H and H' are both true, while $H \vee H'$ is the hypothesis that either H or H' or both are true.

The three analytic plans make reference to four P-values. These are: P_{M1} testing the null hypothesis H_1 of no main effect of factor 1, P_{M2} testing the null hypothesis H_2 of no main effect of factor 2, P_I testing the null hypothesis H_I of no interaction of factors 1 and 2, and $P_{M,1\wedge 2}$ testing the hypothesis $H_1 \wedge H_2$ of no main effect for both factors 1 and 2. Method II does not examine P_I, so it cannot detect an interaction by rejecting H_I. In the simplest case of a balanced 2×2 factorial with Gaussian errors, P_{M1}, P_{M2}, and P_I might be derived from two-sided t-tests on single degree of freedom contrasts, and $P_{M,1\wedge 2}$ might be derived from an F-test combining the two-degrees of freedom for the two main effects. In an unbalanced Gaussian design, these are specific linear hypotheses in a linear model with a constant term, two main effects and one interaction; in particular $P_{M,1\wedge 2}$ is from an F-test about two parameters in this linear model. More generally, these P-values might instead come from likelihood ratio tests under some model or from nonparametric tests (e.g, Patel and Hoel 1973). What is required of the four tests is simply that each test yields a valid P-value; that is, when its null hypothesis is true, the P-value is $\leq \alpha$ with probability at most α for all $0 < \alpha < 1$.

I. Bonferroni-Holm test of main effects and interactions. Sort P_{M1}, P_{M2}, and P_I into nondecreasing order as $P_{(1)} \leq P_{(2)} \leq P_{(3)}$, so for instance $P_{(1)} = \min(P_{M1}, P_{M2}, P_I)$. If $P_{(1)} > \alpha/3$, no hypothesis is rejected and testing stops. If $P_{(1)} \leq \alpha/3$, then the hypothesis of no effect, $H_1 \wedge H_2 \wedge H_I$, is rejected, as is the hypothesis associated with $P_{(1)}$ (one of H_1, H_2 and H_I), and testing continues. If $P_{(1)} \leq \alpha/3$ and $P_{(2)} \leq \alpha/2$, then the hypothesis associated with $P_{(2)}$ is also rejected, and testing continues; otherwise testing stops. If $P_{(1)} \leq \alpha/3$ and $P_{(2)} \leq \alpha/2$ and $P_{(3)} \leq \alpha$, then all three hypotheses about effects (all of H_1, H_2 and H_I) are rejected, so the hypothesis $H_1 \vee H_2 \vee H_I$ is rejected.

II. Bonferroni-Holm test of main effects only. Two P-values are computed, P_{M1} for the main effect of factor 1, P_{M2} for the main effect of factor 2. If $\min(P_{M1}, P_{M2}) > \alpha/2$, no hypothesis is rejected and testing stops. If $\min(P_{M1}, P_{M2}) \leq \alpha/2$, then the hypothesis of no main effects, $H_1 \wedge H_2$, is rejected, as is the hypothesis associated with $\min(P_{M1}, P_{M2})$ (either H_1 or H_2), and testing continues. If $\min(P_{M1}, P_{M2}) \leq \alpha/2$ and $\max(P_{M1}, P_{M2}) \leq \alpha$, then the hypothesis associated with $\max(P_{M1}, P_{M2})$ is also rejected (either H_1 or H_2) so $H_1 \vee H_2$ is rejected. The interaction is not tested.

III. Structured testing, main effects first, then interaction. A single P-value, $P_{M,1\wedge 2}$ is computed to test the null hypothesis of no main effects. If $P_{M,1\wedge 2} > \alpha$, no hypothesis is rejected and testing stops. If $P_{M,1\wedge 2} \leq \alpha$, the hypothesis of no main effects, $H_1 \wedge H_2$, is rejected, and testing continues. If $P_{M,1\wedge 2} \leq \alpha$ and $P_{M1} \leq \alpha$, then the hypothesis H_1 of no main effect of factor 1 is rejected, and also if $P_{M,1\wedge 2} \leq \alpha$ and $P_{M2} \leq \alpha$, then the hypothesis H_2 of no main effect of factor 2 is rejected. If either $P_{M1} > \alpha$ or $P_{M2} > \alpha$, testing stops. Otherwise, if $P_{M,1\wedge 2} \leq \alpha$ and $P_{M1} \leq \alpha$ and $P_{M2} \leq \alpha$, then $H_1 \vee H_2$ is rejected, and the interaction is tested, whereupon if also $P_I \leq \alpha$, then the hypothesis H_I of no interaction is also rejected. If $P_{M,1\wedge 2} \leq \alpha$ and either $P_{M1} \leq \alpha$ or $P_{M2} \leq \alpha$ then at least one main effect has been identified, and the probability of this is labeled identify in Table 1.

Although our main interest is quantitative comparisons of the power of these three procedures, one qualitative comparison provides some insight. The hypothesis $H_1 \vee H_2$ says at least one of the two treatments has no main effect. If one were testing $H_1 \vee H_2$ alone using intersection-union testing, then $H_1 \vee H_2$ would be rejected at level α if both $P_{M1} \leq \alpha$ and $P_{M2} \leq \alpha$; see Berger (1982) and also Lehmann (1952). By contrast, procedure III rejects $H_1 \vee H_2$ if $P_{M,1\wedge 2} \leq \alpha$ and $P_{M1} \leq \alpha$ and $P_{M2} \leq \alpha$; however, in a large balanced factorial under the usual Gaussian model, $P_{M1} \leq \alpha$ and $P_{M2} \leq \alpha$ implies $P_{M,1\wedge 2} \leq \alpha$ (see Miller 1981, §3.7, Figure 12), so in this case procedure III rejects $H_1 \vee H_2$ whenever intersection-union testing rejects $H_1 \vee H_2$. In contrast, procedure II rejects $H_1 \vee H_2$ if $\min(P_{M1}, P_{M2}) \leq \alpha/2$ and $\max(P_{M1}, P_{M2}) \leq \alpha$, so procedure II may fail to reject $H_1 \vee H_2$ when intersection-union testing would reject it. Procedure I rejects $H_1 \vee H_2$ if $\min(P_{M1}, P_{M2}, P_I) \leq \alpha/3$, $\min(P_{M1}, P_{M2}) \leq \alpha/2$ and $\max(P_{M1}, P_{M2}) \leq \alpha$, so procedure I may fail to reject $H_1 \vee H_2$ when procedure II would reject it, and it may fail to reject $H_1 \vee H_2$ when procedure II would fail to reject it but intersection-union testing would reject it. Albeit limited in scope, this qualitative comparison favors Procedure III.

The Bonferroni-Holm procedures require a P-value smaller than α/K if K hypotheses are tested, whereas Plan III rejects hypotheses when appropriate P-values are less than α. There are several ways to see that Plan III controls the probability π of at least one false rejection. Here are two ways.

Consideration of cases. There are three hypotheses H_1, H_2, and H_I, each of which may be true or false, making $2^3 = 8$ possible cases. In an elementary if slightly tedious manner, these eight cases may be considered one at a time to verify that, in each case, procedure III has $\pi \leq \alpha$. To illustrate, consider two of the eight cases. If $H_1 \wedge H_2 \wedge H_I$ is true, a false rejection of at least one true null hypothesis occurs if and only if $P_{M,1\wedge 2} \leq \alpha$, but in this case this happens with probability at most α. If only H_2 is true, then H_1 and H_I are both false and hence cannot be falsely rejected, so a false rejection occurs if and only if $P_{M1} \leq \alpha$ which in this case occurs with probability at most α. And so on.

Sequentially exclusive partition of hypotheses. The ordered sequence of three sets of hypotheses $\langle \{H_1 \wedge H_2\}, \{H_1, H_2\}, \{H_I\} \rangle$ has the property that at most one hypothesis in a set is true if all the hypotheses in earlier sets are false. This is trivially true of $\{H_1 \wedge H_2\}$ and $\{H_I\}$ because these two sets contain only one hypothesis, so they contain at most one true hypothesis. Now, $\{H_1, H_2\}$ might contain two true hypotheses, but $\{H_1, H_2\}$ cannot contain two true hypotheses if $H_1 \wedge H_2$ is false. From this structure alone — known as a sequentially exclusive partition of hypotheses — it follows immediately from Proposition 3 in Rosenbaum (2008) that the probability of at least one false rejection in procedure III is at most α. The partition just mentioned had an ordered sequence of three sets of hypotheses, with 1, 2 and 1 hypotheses in the three consecutive sets. The same reasoning works for an infinite totally ordered collection of sets of hypotheses where each set of hypotheses may contain infinitely many hypotheses; see Rosenbaum (2008; 2010, §19). For instance, either the collection or the sets within the collection may be indexed by a real parameter.

Methods I, II and III are intended to provide a contrast between procedures built from the Bonferroni inequality which equitably subdivide α and structured testing procedures that organize testing termination without subdividing α. It is easy to build procedures which blur this distinction, as is usefully done in several of the references; e.g., indeed, this is already true of Holm's (1979) procedure. For instance, one could produce a hybrid which first does method II, and if $H_1 \vee H_2$ is rejected because $\min(P_{M1}, P_{M2}) \leq \alpha/2$ and $\max(P_{M1}, P_{M2}) \leq \alpha$, goes on to reject H_I if $P_I \leq \alpha$, so there is both splitting of α in testing main effects and a termination structure without splitting α in testing interactions. Our current purpose, however, is to contrast the power of a few quite distinct procedures, rather than introduce many shades of grey.

As with many other descendents of the closed testing of Marcus, Peritz and Gabriel (1976), method III makes use of ideas found in closed testing but is not itself an instance of closed testing. In closed testing in its original form, if one wished to test both main effects and their interaction, one would first test $H_1 \wedge H_2 \wedge H_I$ at level α, stopping if this hypothesis was not rejected. If $H_1 \wedge H_2 \wedge H_I$ is rejected at level α, one would then test at level α three more intersection hypotheses, namely $H_1 \wedge H_2$, $H_1 \wedge H_I$ and $H_2 \wedge H_I$. If both $H_1 \wedge H_2$ and $H_1 \wedge H_I$ were rejected at level α, then H_1 would be tested, again at level α. In closed testing in its original form, the investigator may never reach the stage where main effects are tested separately from the interaction. In contrast, in method III, the investigator tests main effects first without reference to the interaction, but nonetheless may test for interaction when two main effects are discovered. As Holm (1979) observed, Holm's procedure is an instance of closed testing implemented using the Bonferroni inequality as the basis for testing the intersection of several hypotheses. The focus of the current paper is a comparison of the power of three procedures that test main effects immediately.

The three methods can be applied also to an $R \times C$ two-factor factorial design with $R \geq 2$ and $C \geq 2$, for instance by using suitable F-tests in a Gaussian linear model. We do not consider $R > 2$ and $C > 2$ because the focus in the current paper is on the power of the three procedures against various alternatives, and it is convenient that these alternatives for the 2×2 factorial may be described in terms of just three parameters. Although one can devise structured testing approaches for more than two factors, method III as described is not applicable with more than two factors.

3 Power of the three analytic plans

Table 1 gives the powers of the three analytic plans, I, II, and III in §2 to reject various hypotheses for eight possible treatment effects. The eight possible effects, A-H, appear at the top of Table 1. For instance, in setting A, each factor has a main effect of size 0.5, and there is no interaction, so if both factors are applied at their high levels, the effect is $1 = 0.5 + 0.5$ when compared to the low-low group. The power is computed for Gaussian errors with known standard deviation one and ten observations per group. For some details of the computation, see the Appendix. (As is familiar with power calculations for the Normal distribution, it is not the sample size, the standard deviation or the treatment effects that determine the power, but rather a noncentrality parameter that summarizes these quantities. For instance, if the effects and the standard deviation were both doubled, the powers would be the same.)

The hypotheses tested by methods I, II, and III in §2 are not quite the same, and the methods terminate when different events occur. For instance, by definition: (a) hypothesis $H_1 \wedge H_2 \wedge H_I$ is rejected in method I if any of H_1, H_2, H_I is rejected — that is, if $\min(P_{M1}, P_{M2}, P_I) \leq \alpha/3$; (b) hypothesis $H_1 \wedge H_2$ is rejected by method II if either H_1 or H_2 is rejected — that is, if $\min(P_{M1}, P_{M2}) \leq \alpha/2$; whereas, $H_1 \wedge H_2$ is rejected by method III if $P_{M,1\wedge 2} \leq \alpha$. In particular, even in an infinitely large sample, method II might correctly conclude that both main effects are present by rejecting $H_1 \vee H_2$, but this correct conclusion might fail to give an adequate description because a substantial but untested interaction is also present. In this sense, methods I, II and III are running somewhat different risks to test somewhat different hypotheses. With that caution firmly in mind, we turn to an examination of power.

The chance of rejecting at least one hypothesis is much higher for structured testing. For instance, in situation A, method III rejects $H_1 \wedge H_2$ with probability 0.50, method II rejects $H_1 \wedge H_2$ with probability 0.44, and method I rejects $H_1 \wedge H_2 \wedge H_I$ with probability 0.38; otherwise, these methods reject no hypothesis. When structured testing rejects the hypothesis of no main effect, it typically identifies a specific effect; see identify in Table 1. Because method I gives equal emphasis to main effects and interactions, it generally has lower power than methods II and III to detect main effects.

In case E, there are two substantial main effects and an interaction of the same magnitude. Methods II and III operate under the premise that detecting main effects is more important than detecting interactions, whereas Method I gives equal emphasis to main effects and interactions. Method I is at its best and method II is at its worst in case E, because method I has an 86% chance of detecting each effect, while method II cannot detect interactions. In case E, Method III has an 99% chance of rejecting the hypothesis of no main effects, a 98% chance of identifying at least one main effect, an 88% chance of detecting each main effect, and a 69% chance of detecting the interaction.

In case C, there are two main effects and a smaller interaction. None of the procedures has much chance of detecting the interaction: the power is zero for method II, and is low for methods I and III. Nonetheless, structured testing has the highest power to detect the main effects. The situation is similar in case F. In case D, method I is more likely to detect the interaction than method III, but method III has more power to detect main effects.

In case G, only one factor has an effect, and the three methods exhibit similar performance. In case H, one factor has a larger effect than the other, and structured testing has slightly better power than method II.

If higher responses are better responses and if both main effects are positive, then from a clinician's point of view there is a marked asymmetry between failing to detect a positive and a negative interaction. A negative interaction might be a reason for avoiding joint use of the two treatments. Table 1 has both positive and negative interactions. The power of two-sided tests in balanced designs is, however, symmetrical in the sign of the interaction. Indeed, this is also true of the signs of the main effects. That is, if one erased the first four rows of Table 1, keeping the main effects and interactions, and if one changed the signs of the main effects or the interactions, then the powers in the bottom of Table 1 would be unchanged.

Table 1: Power of three analytic plans to reach various conclusions with eight possible patterns of treatment effects. In situations A and B, the two factors have effects that are additive without interaction. In situations C, E and F, the simultaneous application of both factors has an effect greater than the sum of their separate effects. In situation D, the simultaneous application of both factors has an effect less than the sum of their separate effects. In situation G, only factor 1 has a main effect. In situation H, factor 1 has a larger main effect than factor 2. The event identify occurs if $P_{M,1\wedge 2} \le \alpha$ and either $P_{M1} \le \alpha$ or $P_{M2} \le \alpha$ signifying that method III has identified a specific main effect.

A 2 × 2 Factorial with 8 Possible Treatment Effects, A-H

		Mean Response at 2 × 2 Factor Levels							
Factor 1 Level	Factor 2 Level	A	B	C	D	E	F	G	H
Low	Low	0	0	0	0	0	0	0	0
High	Low	.5	.75	.5	1	0	.45	1	.75
Low	High	.5	.75	.5	1	0	.45	0	.25
High	High	1	1.5	1.5	1	2	1.8	1	1
	Main Effect 1	.5	.75	.75	.5	1	.9	1	.75
	Main Effect 2	.5	.75	.75	.5	1	.9	0	.25
	Interaction	0	0	.25	-0.5	1	.45	0	0

Power of Three Analytic Plans to Reach Various Conclusions

Analytic Plan	Null Hypothesis	Probability of Rejection							
I. Holm's Method Testing Main Effects and Interaction	H_1	.22	.52	.53	.23	.86	.73	.78	.49
	H_2	.22	.52	.53	.23	.86	.73	.02	.07
	H_I	.02	.03	.08	.23	.86	.25	.02	.02
	$H_1 \wedge H_2 \wedge H_I$	.38	.75	.76	.50	.99	.91	.79	.53
II. Holm's Method Testing Just Main Effects	H_1	.28	.61	.61	.28	.87	.79	.82	.56
	H_2	.28	.61	.61	.28	.87	.79	.05	.10
	$H_1 \wedge H_2$	.44	.80	.80	.44	.97	.93	.83	.58
III. Structured Testing	$H_1 \wedge H_2$	.50	.86	.86	.50	.99	.96	.82	.60
	identify	.49	.85	.85	.49	.98	.95	.81	.59
	H_1	.31	.64	.64	.31	.88	.81	.81	.56
	H_2	.31	.64	.64	.31	.88	.81	.05	.12
	H_I	.01	.02	.05	.04	.69	.20	.00	.00

The calculations in Table 1 are exact but assume that the variance is known. Using simulation, we calculated the powers with an unknown variance and either 8 or 10 subjects per treatment group, which is a small number for a clinical trial, and the results were qualitatively very similar to Table 1 and so are not reported.

4 Summary

In the 2×2 factorial design, an investigator may give priority to main effects, while wishing to look for interactions if main effects are found. In the varied situations in Table 1, structured testing yielded an increase in power for a given sample size, or a reduction in sample size for a given power, when compared to the most commonly used procedures that also control the probability of false rejections in multiple tests.

The Appendix presents a formula for power or sample size calculations when the degrees of freedom for error are sufficiently large that they have little effect on power. Software for sample size calculations in small samples is available from the first author.

Appendix: example of power calculations

We illustrate one of the less standard power computations in Table 1. In each situation in Table 1, the two main effects contrasts, say $\widehat{\theta}_1$ and $\widehat{\theta}_2$, are independent Normal random variables, $\widehat{\theta}_j \sim N\left(\theta_j, \sigma_j^2\right)$, where σ_j^2 is known, so $\widehat{\theta}_j^2/\sigma_j^2$ has a chi-square distribution with one degree of freedom and noncentrality parameter $\omega_j = \theta_j^2/\sigma_j^2$; write $f\left(\cdot\,;\omega_j\right)$ for this density and $F\left(\cdot\,;\omega_j\right)$ for the cumulative distribution. Let a be the upper α percentile of the central chi-square distribution on one degree of freedom, and let b be the upper α quantile of the central chi-square distribution on two degrees of freedom. For $\alpha = 0.05$, the constants are $a = 3.84$ and $b = 5.99$. In Table 1, hypothesis H_1 is rejected if $\widehat{\theta}_1^2/\sigma_1^2 + \widehat{\theta}_2^2/\sigma_2^2 \geq b$ and $\widehat{\theta}_1^2/\sigma_1^2 \geq a$, and this can happen in two mutually exclusive ways: (i) $\widehat{\theta}_1^2/\sigma_1^2 \geq b$ which happens with probability $1 - F\left(b\,;\omega_1\right)$, or (ii) for some x, $a \leq x < b$, $\widehat{\theta}_1^2/\sigma_1^2 = x$ and $\widehat{\theta}_2^2/\sigma_2^2 \geq b - x$, so the chance that H_1 is rejected is

$$\beta\left(\omega_1, \omega_2\right) = \int_a^b f\left(x; \omega_1\right)\left\{1 - F\left(b - x\,; \omega_2\right)\right\} dx + \left\{1 - F\left(b\,; \omega_1\right)\right\}.$$

In Table 1, $\sigma_j^2 = 1/10$, $j = 1, 2$. In situation C in Table 1, $\theta_1 = \theta_2 = 0.75$, so $\omega_1 = \omega_2 = 0.75^2/\left(1/10\right) = 5.625$, and $\beta\left(5.625, 5.625\right) = 0.64176$, as in Table 1. The R function tiopower computes $\beta\left(\mathrm{ncp}_1, \mathrm{ncp}_2\right)$ for given α.

```
> tiopower
function(ncp1,ncp2,alpha=0.05){
 q12<-qchisq(1-alpha,2)
 q1<-qchisq(1-alpha,1)
 g<-function(x){dchisq(x,1,ncp=ncp1)*(1-pchisq(q12-x,1,ncp=ncp2))}
 integrate(g,q1,q12)$value+1-pchisq(q12,1,ncp1)
}


> tiopower((.75^2)/.1,(.75^2)/.1)
[1] 0.6417632
```

References

[1] Bauer, P. (1997), “A note on multiple testing in dose-finding,” *Biometrics*, 53, 1125-1128.

[2] Bauer, P. and Kieser, M. (1996), “A unifying approach for confidence intervals and testing of equivalence and difference,” *Biometrika*, 83, 934-7.

[3] Berger, R. L. (1982), “Multiparameter hypothesis testing and acceptance sampling,” *Technometrics* 24, 295-300.

[4] Brown, B. G., Zhao, X. Q., Chait, A., Fisher, L. D., Cheung, M. C., Morse, J. S., Dowdy, A. A., Marino, E. K., Bolson, E. L., Alaupovic, P., Frohlich, J, Albers, J. J. (2001), “Simvastatin and niacin, antioxidant vitamins, or the combination for the prevention of coronary disease,” *New England Journal of Medicine,* 345, 1583-92.

[5] Byar, D. P. and Piantadosi, S. (1985), “Factorial designs for randomized clinical trials,” *Cancer Treatment Reports*, 69, 1055-63.

[6] Ehm, W., Kornmeier, J. and Heinrich, S. P. (2010), “Multiple testing along a tree,” *Electronic Journal of Statistics*, 4, 461-471.

[7] Ellenberg, S. S., Finkelstein, D. M., and Schoenfeld, D. A. (1992), “Statistical issues arising in AIDS clinical trials,” *Journal of the American Statistical Association*, 87, 562-569.

[8] Finner, H. and Strassburger, K. (2002), “The partitioning principle: a powerful tool in multiple decision theory,” *Annals of Statistics*, 30, 1194-1213.

[9] Finos, L. and Farcomeni, A. (2010), “k-FWER control without p-value adjustment,

with application to detection of genetic determinants of multiple sclerosis in Italian twins," *Biometrics*, to appear.

[10] Goeman, J. J., Solari, A. (2010), "The sequential rejection principle of familywise error control," *Annals of Statistics*, 38, 3782-3810.

[11] Hochberg, Y., Tamhane, A.C. (1987), *Multiple Comparison Procedures*, New York: Wiley.

[12] Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

[13] Hommel, G., Kropf, S. (2005), "Tests for differentiation in gene expression using a data-driven order or weights for hypotheses," *Biometrical Journal*, 47, 554-562.

[14] Hommel, G., Bretz, F. and Mauer, W. (2007), "Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies," *Statistics in Medicine*, 26, 4063-4073.

[15] Hsu, J. C. and Berger, R. L. (1999), "Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies," *Journal of the American Statistical Association,* 94, 468-75.

[16] Koch, G.G. and Gansky, S.A. (1996), "Statistical considerations for multiplicity in confirmatory protocols," *Drug Information Journal*, 30, 523-33.

[17] Lehmann, E. L. (1952), "Testing multiparameter hypotheses," *Annals of Mathematical Statistics* 23, 541-52.

[18] Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York: Springer.

[19] Marcus, R., Peritz, E., Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63, 655-60.

[20] Miller, R. G., Jr. (1981), *Simultaneous Statistical Inference* (2^{nd} edition), New York: Springer.

[21] Meinshausen, N. (2008), "Hierarchical testing of variable importance," *Biometrika*, 95, 665-678.

[22] Papp, K. V., Walsh, S. J. and Snyder, P. J. (2009), "Immediate and delayed effects of cognitive interventions in healthy elderly: A review of current literature and future directions," *Alzheimer's & Dementia*, 5, 50-60.

[23] Patel, K. M. and Hoel, D. G. (1973), "A nonparametric test for interaction in factorial experiments," *Journal of the American Statistical Association*, 68, 615-620.

[24] Rosenbaum, P. R. (2008), "Testing hypotheses in order," *Biometrika*, 95, 248-252.

[25] Rosenbaum, P. R., Silber, J. H. (2009), "Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units," *Journal of the American Statistical Association*, 104, 501-511.

[26] Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer.

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

[28] Stampfer, M. J., Buring, J. E., Willett, W., Rosner, B., Eberlein, K. andHennekens, C. H. (1985), "The 2 × 2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians," *Statistics in Medicine*, 4, 111-116.

[29] Stone, G. W., Grines, C. L., Cox, D. A., Garcia, E., Tcheng, J. E., Griffin, J. J., Guagliumi, G., Stuckey, T., Turco, M., Carroll, J. D., Rutherford, B. D., Lansky, A. J. and the Controlled Abciximab and Device Investigation to Lower Late Angioplasty Complications (CADILLAC) Investigators (2002), "Comparison of angioplasty with stenting, with or without abciximab, in acute myocardial infarction," *New England Journal of Medicine,* 346, 957-66.

[30] Westfall, P. H. and Krishen, A. (2001), "Optimally weighted, fixed sequence and gatekeeper multiple testing procedures," *Journal of Statistical Planning and Inference*, 99, 25-40.

[31] Willis, S.L., Tennstedt, S.L., Marsiske, M., et al. (2006), "Long-term effect of cognitive training on everyday functional outcomes in older adults," *Journal of the American Medical Association,* 296, 2805-2814.