Statistica Sinica **16**(2006), 933-951

AN INVESTIGATION OF TWO-STAGE TESTS

Marc Vandemeulebroecke

Otto-von-Guericke-Universität Magdeburg and Schering AG

Abstract: Two-stage tests may be defined in terms of a combination function for the p-values of the separate stages, or alternatively by specifying a conditional error function, i.e., the conditional probability for an erroneous rejection given the first stage. Examples have been published suggesting that these two approaches are essentially equivalent. We provide a formal link between them that yields a general framework for two-stage tests. Our viewpoint leads to an overall p-value notion that covers different previously proposed concepts, and it allows an easy construction of new two-stage tests. One particular test is further characterized.

Key words and phrases: Adaptive design, combination test, conditional error function, interim analysis, overall p-value, two-stage test.

1. Introduction

Armitage, McPherson and Rowe (1969) quantified the inflation of the type I error in (naive) sequential testing. Since then numerous statistical procedures have been proposed that adjust for this inflation, especially in the field of clinical trials, seeking to maximize flexibility in trial conduct and to minimize patient exposure or costs. The group sequential tests of Pocock (1977) and O'Brien and Fleming (1979) were more practical than the purely sequential, in theory optimal, Sequential Probability Ratio Test of Wald (1945) (see also Wald and Wolfowitz (1948)). Group sequential variants were provided by DeMets and Ware (1980, 1982), Gould and Pecore (1982) and others. Wang and Tsiatis (1987) and Pampallona and Tsiatis (1994) generalized these procedures to the "Δ-class" of group sequential boundaries. Lan and DeMets (1983) introduced more flexibility with the alpha-spending function approach, setting aside the need for prespecified stage sizes and allowing an arbitrary apportionment of the type I error over the stages. Bauer (1989a) put forward the principle of combining the p-values from separate stages. This principle permits a wide range of data-driven design modifications, including an adaptive choice of the sample size, the test statistic or even the null hypothesis. Two prominent examples apply combination methods originally conceived for meta-analyses to the context of adaptive trials. Bauer and Köhne (1994) suggested

the use of Fisher's product criterion (Fisher (1932), see also Bauer (1989a, 1989b)). Lehmacher and Wassmer (1999) put emphasis on the inverse normal method (Mosteller and Bush (1954), see also Bauer and Köhne (1994) and Cui, Hung and Wang (1999)). Motivated by the wish to extend a study in order to reach a certain *conditional power* (Lan, Simon and Halperin (1982)), Proschan and Hunsberger (1995) proposed to define a two-stage test by a *conditional error function*. This function specifies the conditional probability for an erroneous rejection of the null hypothesis given the first stage. More recent works include the "self-designing trials" of Fisher (1998) and Shen and Fisher (1999); modifications of the Proschan-Hunsberger procedure by Liu and Chi (2001) and Li, Shih, Xie and Lu (2002); adaptive multistage designs by Müller and Schäfer (2001) and Brannath, Posch and Bauer (2002); and lately, an interesting geometrical characterization of two-stage tests by Proschan (2003).

All these approaches are interrelated to a greater or lesser extent, as has been pointed out by Posch and Bauer (1999), Wassmer (1999, 2000), Bauer, Brannath and Posch (2001) and Jennison and Turnbull (2003) among others. In the context of trials with only two stages, Posch and Bauer (1999) and Wassmer (1999, 2000) presented examples that suggest a general correspondence between a conditional error function and a function that combines two p-values in the spirit of Bauer (1989a). The idea of this general correspondence is commonly accepted, but its nature has not yet been thoroughly explored. This is the first purpose of the present article. We show that a p-value combination function corresponds in fact to a whole family of conditional error functions, specifying level-α-tests for each level α between 0 and 1. Based on this link, our second focus lies in defining overall p-values for two-stage tests. Overall p-values are of interest as a "measure of certainty" and may be used to construct multistage tests by recursive combination as presented by Brannath, Posch and Bauer (2002). Previous definitions based on a p-value combination function (Brannath, Posch and Bauer (2002)) or a family of conditional error functions (Liu and Chi (2001)) fit in our framework. Finally, and this is our third point, we illustrate that our concept allows an easy construction of new two-stage tests and a more flexible handling of known two-stage tests.

We hope to contribute to the understanding of two-stage tests by providing a general, formally rigorous and geometrically intuitive framework that covers several previously proposed concepts. The rest of this paper is organized as follows. Section 2 recapitulates the p-value combination function approach and the conditional error function approach. A formal link between these approaches is provided in Section 3. Overall p-values are the subject of Section 4, followed by an example in Section 5. Section 6 completes the article with a brief discussion.

Technical details and key steps of proof are provided in three appendices. To avoid problems with conflicting directional decisions, we assume one-sided testing throughout. Integrals are taken over the unit interval unless otherwise specified.

2. Two Approaches

A two-stage procedure for testing a null hypothesis H_0 may be defined in terms of the overall level α, stopping bounds α_1 and α_0, a parameter α_2 and a function $C(\cdot,\cdot)$ to combine the p-values of the two stages. The bound α_1 is the local level of the test based on the first stage. The parameter α_2 is the local level of the test based on both stages, ignoring the two-stage nature of the design. The quantities α, α_0 and α_1 are subject to the condition $0 \le \alpha_1 \le \alpha \le \alpha_0 \le 1$, and C must have some regularity properties that will be defined later. After computing the p-value p_1 of the first stage, the test stops with rejection of H_0 if $p_1 \le \alpha_1$, and it stops without rejection of H_0 ("for futility") if $p_1 > \alpha_0$. If $\alpha_1 < p_1 \le \alpha_0$, the test proceeds, the p-value p_2 of the second stage is computed, and the "combination test" is carried out. H_0 is then rejected if and only if $C(p_1, p_2)$ is not greater than some threshold $c(\alpha_2)$ that is determined by the local level α_2 of this combination test. Given C, the choice of α_0, α_1 and $c(\alpha_2)$ is constrained by the desired overall level α for the two-stage procedure, assuming that p_1 and p_2 are independent and uniformly distributed on $[0,1]$ under H_0. As Bauer (1989a) pointed out, however, this assumption is not necessary for the level α to be kept. It suffices that under H_0 the distribution of p_1 and the conditional distribution of p_2 given p_1 are stochastically not smaller than the uniform distribution on $[0,1]$. Brannath, Posch and Bauer (2002) called this property *p-clud*. The rejection region of such a test can be visualized as the area

$$\{p_1 \le \alpha_1\} \cup (\{C(p_1,p_2) \le c(\alpha_2)\} \cap \{p_1 \le \alpha_0\})$$

in the unit square. Choosing $\alpha_0 = \alpha = \alpha_1$ yields a single stage test as a special case.

This idea of combining the p-values from separate stages was put forward by Bauer (1989a). As an example, Bauer and Köhne (1994) and Bauer and Röhmel (1995) proposed the choice of $C(p_1,p_2) = p_1 p_2$, yielding $c(\alpha_2) = \exp(-\chi^2_{4,\alpha_2}/2)$ (Fisher (1932)) which is traditionally written as c_{α_2}. By χ^2_{4,α_2} we denote the $(1-\alpha_2)$-quantile of the central χ^2-distribution with 4 degrees of freedom. Supposing $c_{\alpha_2} \le \alpha_1$, the parameters need to satisfy $\alpha_1 + c_{\alpha_2}(\ln\alpha_0 - \ln\alpha_1) = \alpha$ for an overall test level of α. We refer to this test as Fisher's combination test.

An alternative approach, originally posed by Proschan and Hunsberger (1995), defines a two-stage test by specifying its conditional error function, i.e., the conditional probability for an erroneous rejection given the first stage. Following Wassmer (1999, 2000), we write this function as a function in p_1. Any

nonincreasing function $\bar{\alpha}$ with values in $[0,1]$ may be used. For technical reasons we additionally assume $\bar{\alpha}$ to be left continuous. In this approach the same testing procedure is implemented as in the p-value combination function approach, with the combination test criterion $C(p_1,p_2) \le c(\alpha_2)$ replaced by $p_2 \le \bar{\alpha}(p_1)$. That is, α, α_0 and α_1 are chosen to satisfy $0 \le \alpha_1 \le \alpha \le \alpha_0 \le 1$. If $p_1 \le \alpha_1$ or $p_1 > \alpha_0$, the test stops after the first stage, with or without rejection of H_0, respectively. Otherwise, H_0 is rejected if and only if $p_2 \le \bar{\alpha}(p_1)$. We can set $\alpha_2 = \int \bar{\alpha}(p_1)\,dp_1$, and the interpretation of all parameters is the same as in the p-value combination function approach. The choice of the parameters is constrained by $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}(p_1)\,dp_1 = \alpha$, and the rejection region in the unit square is now

$$\{p_1 \le \alpha_1\} \cup (\{p_2 \le \bar{\alpha}(p_1)\} \cap \{p_1 \le \alpha_0\}).$$

Again, the constraining equation makes use of the assumption that p_1 and p_2 are independent and uniformly distributed on $[0,1]$ under H_0. The procedure still keeps the level α if p_1 and p_2 are p-clud under H_0. Note that α_1 and α_0 are imposed on $\bar{\alpha}$ rather than being part of it. In a common alternative notation, $\max\{p_1;\, \bar{\alpha}(p_1) = 1\}$ and $\inf\{p_1;\, \bar{\alpha}(p_1) = 0\}$ take over the roles of α_1 and α_0, respectively, and the choice of $\bar{\alpha}$ is constrained by $\int \bar{\alpha}(p_1)\,dp_1 = \alpha$. In our view, however, this obscures the analogy between the two approaches.

For example, $\bar{\alpha}(p_1) = \min\{1, c_{\alpha_2}/p_1\}$ is a conditional error function that depends on α_2. Together with α, α_0 and α_1, constrained by $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}(p_1)\,dp_1 = \alpha_1 + c_{\alpha_2}(\ln\alpha_0 - \ln\alpha_1) = \alpha$ (and $c_{\alpha_2} \le \alpha_1$), it specifies a two-stage test.

Clearly, this test is identical to Fisher's combination test. This and other examples, first presented by Posch and Bauer (1999) and Wassmer (1999, 2000), have led to the common understanding that the two approaches complement one another. The next section investigates their relationship in a formal way in a general setting.

3. A Formal Framework for Two-stage Tests

To motivate the definitions that follow, imagine p_1 and p_2 are combined by some function C that is continuous and strictly increasing in both arguments. C defines a "rising surface" over the unit square, and the null hypothesis is rejected if $C(p_1,p_2)$ does not exceed some prespecified height $H = c(\alpha_2)$ (early stopping not considered). The level curve $\{C(p_1,p_2) = H\}$ may be thought of as the boundary of the rejection region $\{C(p_1,p_2) \le H\}$. We can write this level curve as a function $\bar{\alpha}^H$ in p_1. For a given value p_1, the null hypothesis is then rejected if $p_2 \le \bar{\alpha}^H(p_1)$, due to the monotonicity properties of C. Figure 1 illustrates this idea for Fisher's combination test. Assuming p_2 is uniformly distributed on $[0,1]$, the conditional rejection probability given p_1 is $\Pr(p_2 \le \bar{\alpha}^H(p_1)) = \bar{\alpha}^H(p_1)$. Thus, $\bar{\alpha}^H$ is the conditional error function.

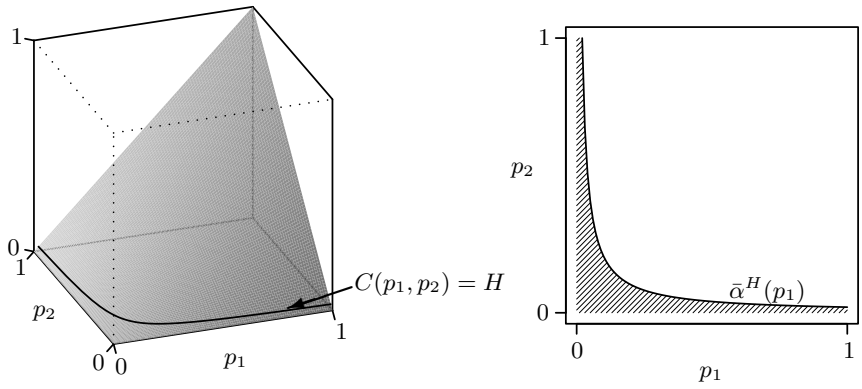

Figure 1. Fisher's combination test at the level 0.1 without early stopping. Left: surface of $C(p_1, p_2) = p_1 p_2$ over the unit square, with the level curve at height $H = c_{0.1} = 0.0205$. Right: the same level curve as a function $\bar{\alpha}^H$ in p_1, with the rejection region shaded.

However, this reasoning is not always applicable in a straightforward way. In Figure 1, for example, the level curve $\bar{\alpha}^H$ is not defined over the entire unit interval. In a more general situation, C may not be continuous, or it may have constant regions, and the level sets $\{C(p_1, p_2) = H\}$ can have unusual shapes. Conversely, it is not clear how to find a combination function C that "corresponds" to a given conditional error function. A generalization of the level curve idea is needed. For this purpose we now develop the following framework.

We call any function $C : (0,1)^2 \to \mathbb{R}$ a *p-value combination function* if $C(\cdot\,, p_2)$ is nondecreasing and left continuous for all $p_2 \in (0,1)$, and $C(p_1, \cdot\,)$ is nondecreasing for all $p_1 \in (0,1)$. By a *conditional error function* we mean any nonincreasing and left continuous function $\bar{\alpha} : (0,1) \to [0,1]$. The following properties follow by elementary arguments.

Property 1. *For any p-value combination function C, $\bar{\alpha}^H(p_1) = \max\{\sup[p_2 \in (0,1);\ C(p_1, p_2) \le H],\ 0\}$ defines a conditional error function for every $H \in \mathbb{R}$. (We use the convention $\sup(\emptyset) = -\infty$.) If $H \le H'$, then $\bar{\alpha}^H \le \bar{\alpha}^{H'}$ on $(0,1)$. We have $\bar{\alpha}^H \le \bar{\alpha}^{H'}$ on $(0,1)$ if and only if $\int \bar{\alpha}^H(p_1)\,dp_1 \le \int \bar{\alpha}^{H'}(p_1)\,dp_1$, so the family $(\bar{\alpha}^H)_H$ may be reparameterized as $(\bar{\alpha}_h)_h$ such that $h = \int \bar{\alpha}_h(p_1)\,dp_1$. Here, $\bar{\alpha}_h \ne \bar{\alpha}_{h'}$ for any $h \ne h'$. The function $\bar{\alpha}_0$ ($\bar{\alpha}_1$) will exist if and only if C is bounded from below (above); otherwise define $\bar{\alpha}_0 = 0$ ($\bar{\alpha}_1 = 1$) on $(0,1)$. The entire mapping is denoted by $\tilde{\alpha}$, $(\bar{\alpha}_h)_h = \tilde{\alpha}(C)$.*

Property 2. *Let $\mathfrak{a} = (\bar{\alpha}_h)_{h \in [0,1]}$ be a family of conditional error functions satisfying $h = \int \bar{\alpha}_h(p_1)\,dp_1$ for all h, and $\bar{\alpha}_h \le \bar{\alpha}_{h'}$ on $(0,1)$ for any $h \le h'$.*

Then $C(p_1, p_2) = \min\{h \in [0,1];\, \bar{\alpha}_h(p_1) \geq p_2\}$ *defines a p-value combination function* C*. We denote this mapping by* $\tilde{C}$*,* $C = \tilde{C}(\mathfrak{a})$*.*

Property 3. *For any* $\mathfrak{a}$ *as in Property 2,* $\tilde{\alpha}(\tilde{C}(\mathfrak{a})) = \mathfrak{a}$*.*

Note that for any conditional error function $\bar{\alpha}$, $\int \bar{\alpha}(p_1)\, dp_1 = 0$ implies $\bar{\alpha} = 0$ on $(0,1)$, and $\int \bar{\alpha}(p_1)\, dp_1 = 1$ implies $\bar{\alpha} = 1$ on $(0,1)$. The function $\bar{\alpha}^H$ defined in Property 1 can be interpreted as the generalized level curve at height H of the $C(p_1, p_2)$-surface over $(0,1)^2$, possibly completed by the bounds 0 and 1. Details are given in Appendix A. In Property 3, the application of $\tilde{\alpha} \circ \tilde{C}$ first turns $(\bar{\alpha}_h)_{h\in[0,1]}$ into a family $(\bar{\alpha}^H)_{H\in\mathbb{R}}$, with $\bar{\alpha}^H = \bar{\alpha}_H$ for $H \in [0,1]$, $\bar{\alpha}^H = 1$ for $H > 1$, and $\bar{\alpha}^H = 0$ for $H < 0$. By reparameterization, the $\bar{\alpha}^H$ with $H \notin [0,1]$ are cut away again, and the $\bar{\alpha}^H$ with $H \in [0,1]$ are left unchanged. Vice versa, the application of $\tilde{C} \circ \tilde{\alpha}$ to some p-value combination function C would in general compress or stretch the $C(p_1, p_2)$-surface over the unit square vertically. The reparameterization in Property 1 transforms the actual heights H into "standardized heights" h that equal the integral over the respective level curve. In the language of Section 2, H is $c(\alpha_2)$ and h is α_2. However, note that the application of $\tilde{\alpha}$ to C does not necessarily yield conditional error functions $\bar{\alpha}_h$ for every $h \in [0,1]$. For example, a constant function $C(p_1, p_2)$ induces only $\bar{\alpha}_0$ and $\bar{\alpha}_1$. Those C that do induce $\bar{\alpha}_h$ for every $h \in [0,1]$ are called *regular* p-value combination functions.

Based on Properties 1−3, the correspondence between p-value combination functions and families of conditional error functions can be formulated as in Proposition 1.

Proposition 1. *Let* $\mathfrak{A}$ *denote the set of all families* $\mathfrak{a} = (\bar{\alpha}_h)_{h\in[0,1]}$ *of conditional error functions as in Property 2, that is, satisfying* $h = \int \bar{\alpha}_h(p_1)\, dp_1$ *for all* h*, and* $\bar{\alpha}_h \leq \bar{\alpha}_{h'}$ *on* $(0,1)$ *for any* $h \leq h'$*. Then* $\tilde{\alpha}$ *as in Property 1 defines a surjective mapping from the set of all regular p-value combination functions* C *onto* $\mathfrak{A}$*. This mapping reduces to a bijection if we identify any* C*,* C' *with* $\tilde{\alpha}(C) = \tilde{\alpha}(C')$*.*

In simple terms, $\mathfrak{A}$ provides special ways of filling the unit square, conveniently parameterized by the own integral. Figure 2 illustrates this for Fisher's combination test. The p-value combination function $C(p_1, p_2) = p_1 p_2$ induces the family $(\bar{\alpha}^H)_{H\in\mathbb{R}}$ defined by $\bar{\alpha}^H(p_1) = \max\{0, \min\{1, H/p_1\}\}$. It can be reparameterized as $(\bar{\alpha}_{\alpha_2})_{\alpha_2\in[0,1]}$, with $\bar{\alpha}_{\alpha_2}(p_1) = \min\{1, c_{\alpha_2}/p_1\}$ for $0 < \alpha_2 < 1$, $\bar{\alpha}_0 = 0$, and $\bar{\alpha}_1 = 1$. The $\bar{\alpha}_{\alpha_2}$ describe the level curves of the $p_1 p_2$-surface over the unit square, completed by the upper bound 1. The parameter α_2 equals the integral $\int \bar{\alpha}_{\alpha_2}(p_1)\, dp_1$. Note that it is not required that every point in the unit square lies on exactly one of the curves. Infinitely many curves pass through (p_1, p_2) if $p_2 = 1$. Or consider the family specified by $\bar{\alpha}_h = \mathbf{1}_{[0,h]}$, $h \in [0,1]$, where $\mathbf{1}_{[0,h]}(p_1)$

equals 1 if $p_1 \in [0, h]$, and 0 otherwise. Those (p_1, p_2) with $p_2 \in (0, 1)$ lie on none of the curves.

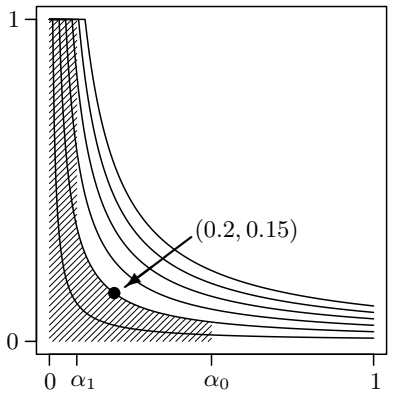

Figure 2. Filling the unit square with the family of level curves $\{C(p_1, p_2) = H\}_H$ where $C(p_1, p_2) = p_1 p_2$ (corresponding to Fisher's combination test). When completed by the upper bound 1, these curves can be written as a family of conditional error functions $(\bar{\alpha}_{\alpha_2})_{\alpha_2}$, with $\bar{\alpha}_{\alpha_2}(p_1) = \min\{1, c_{\alpha_2}/p_1\}$, and $c_{\alpha_2} = \exp(-\chi^2_{4,\alpha_2}/2)$. Supposing $\alpha_1 = 0.0845$ and $\alpha_0 = 0.5$, the overall p-value for $(p_1, p_2) = (0.2, 0.15)$ is 0.1378 (see Section 4). This is the area of the shaded region.

It is important to realize that it does not matter whether the way the unit square is filled stems from a p-value combination function or not. Based on any $\mathfrak{a} = (\bar{\alpha}_h)_{h \in [0,1]} \in \mathfrak{A}$, a two-stage test is implemented as follows.

Method 1. Select α, α_0, α_1 and $\alpha_2 \in [0, 1]$ that satisfy $0 \le \alpha_1 \le \alpha \le \alpha_0 \le 1$ and the level condition $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(p_1)\, dp_1 = \alpha$. Observe the first stage. If $p_1 \le \alpha_1$ or $p_1 > \alpha_0$, then stop with or without rejection of the null hypothesis, respectively. Otherwise conduct the second stage, and reject the null hypothesis if and only if $p_2 \le \bar{\alpha}_{\alpha_2}(p_1)$.

This is the conditional error function approach introduced in the previous section, with $\bar{\alpha} = \bar{\alpha}_{\alpha_2}$. The test keeps the level α if p_1 and p_2 are p-clud under the null hypothesis, and it has exact level α if p_1 and p_2 are independent and uniformly distributed on $[0, 1]$ under the null hypothesis. In particular, a two-stage test with an arbitrarily chosen level $\alpha \in [0, 1]$ can always be constructed. Finally, consider the case that the underlying $\mathfrak{a} \in \mathfrak{A}$ has been induced by a p-value combination function C. Then, by Appendix A(1), we can replace the condition $p_2 \le \bar{\alpha}_{\alpha_2}(p_1)$ in Method 1 by $C(p_1, p_2) \le c(\alpha_2)$, provided that $C(p_1, \cdot\,)$

is left continuous for all p_1. This completes the connection between the two approaches.

Chi and Liu (1999) proposed the same idea of filling the unit square by some suitable family of conditional error functions. They used this concept to design mid-trial sample size re-estimation in case of a misspecified anticipated treatment effect, rather than being motivated by a connection to the p-value combination function approach. The properties required for their conditional error function families differ from those proposed here. In particular, for Fisher's combination test, they concluded that a way their requirements would be "simultaneously satisfied is not readily apparent" (Liu and Chi (2001)). In our framework the corresponding family is easily specified, as sketched after Proposition 1 and illustrated in Figures 1 and 2.

The inverse normal method (Lehmacher and Wassmer (1999), see also Posch and Bauer (1999) and Wassmer (1999, 2000)) serves as another example. In our notation it is represented by

$$C^{\mathrm{INM}}(p_1, p_2) = \frac{\Phi^{-1}(p_1) + \Phi^{-1}(p_2)}{\sqrt{2}},$$

$$\bar{\alpha}_{\alpha_2}^{\mathrm{INM}}(p_1) = \begin{cases} 0 & \text{if } \alpha_2 = 0, \\ \Phi\bigl(\sqrt{2}\,\Phi^{-1}(\alpha_2) - \Phi^{-1}(p_1)\bigr) & \text{if } 0 < \alpha_2 < 1, \\ 1 & \text{if } \alpha_2 = 1, \end{cases}$$

where Φ denotes the distribution function of the standard normal distribution, and $c^{\mathrm{INM}}(\alpha_2) = \Phi^{-1}(\alpha_2)$. Note that, according to Method 1, for a fixed level α the three design parameters α_0, α_1 and α_2 interact due to the level condition, but there is no interdependence of just two of them. For example, we may want to fix α_1 (and α). Then we are still free to manipulate α_0 and α_2. This is not possible in the classical formulation of the inverse normal method, where α_1 and α_2 are directly linked (Li, Shih, Xie and Lu (2002) pointed out the same for their procedure). Thus, Method 1 can not only be used to define new two-stage tests, but also to make known two-stage tests more flexible.

4. Overall p-values for Two-stage Tests

If overall p-values are available for two-stage tests, then multistage tests can be constructed by recursive combination. Brannath, Posch and Bauer (2002) presented this idea. In a two-stage test, the p-value p_2 of the second stage may itself be the overall p-value of another two-stage test. The second stage of this latter test may again be performed in two stages, and so on. Brannath, Posch and Bauer defined an overall p-value function based on the combination

function C. They also noted that other p-value functions can be used, and alluded to a proposal by Liu and Chi (2001) that is based on a family of conditional error functions. Here, the idea of Liu and Chi will be used to define overall p-values within the framework of the previous section; the notion of Brannath, Posch and Bauer may be viewed as a special case. The concept is similar to what Fairbanks and Madsen (1982) proposed in the group sequential setting, and the sample space ordering in the sense of Tsiatis, Rosner and Mehta (1984) is respected.

Lacking a general formal definition, p-values are commonly conceived to represent one of two things (or both): the probability under the null hypothesis of getting observations at least as extreme as the ones actually observed, or the lowest level at which a selected test still rejects the null hypothesis. The latter concept presupposes the availability of a test for each possible level. The former concept requires—in the context of two-stage tests—the specification of what is "at least as extreme" as an observed (p_1, p_2) in the unit square. Is it any (p_1', p_2') with $p_1' + p_2' \leq p_1 + p_2$, or maybe $p_1' p_2' \leq p_1 p_2$? A more general approach would be to fill the unit square by a family of conditional error functions, to select the function that passes through (p_1, p_2), and to call the area below (and including) this function "at least as extreme". In other words, both concepts require an element $(\bar{\alpha}_h)_{h\in[0,1]}$ of $\mathfrak{A}$. On $(\bar{\alpha}_h)_{h\in[0,1]}$ we additionally impose early stopping bounds α_0 and α_1 that are, unlike in Method 1, assumed to be the same for all $\bar{\alpha}_h$. More formally:

Definition 1. For $\mathfrak{a} = (\bar{\alpha}_h)_{h\in[0,1]} \in \mathfrak{A}$ and $0 \leq \alpha_1 \leq \alpha_0 \leq 1$, the function $p : [0,1]^2 \to [0,1]$,

$$p(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0, \\ \alpha_1 + \displaystyle\int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{h^\star}(x)\, dx & \text{otherwise,} \end{cases}$$

with $h^\star = \tilde{C}(\mathfrak{a})(p_1, p_2) = \min\{h \in [0,1];\, \bar{\alpha}_h(p_1) \geq p_2\}$ as defined in Property 2, is called the *overall p-value* (*function*).

Note that if $\alpha_1 = 0$ and $\alpha_0 = 1$, then $p = \tilde{C}(\mathfrak{a})$ on $(0,1)^2$. By Property 3, any $\mathfrak{a} \in \mathfrak{A}$ may thus be interpreted as the family of level curves of its own overall p-value function.

Fisher's combination test, as depicted in Figure 2, again serves as an illustration. Suppose $\alpha_1 = 0.0845$ and $\alpha_0 = 0.5$ (yielding an overall test level of $\alpha = 0.1$ if $\alpha_2 = 0.05$). If $p_1 = 0.2$ and $p_2 = 0.15$, then h for which $\bar{\alpha}_h(p_1) \geq p_2$ barely occurs is $h^\star = \int \bar{\alpha}^{0.03}(x)\, dx$, and the overall p-value is $0.0845 + \int_{0.0845}^{0.5} \bar{\alpha}^{0.03}(x)\, dx = 0.0845 + 0.03\{\ln(0.5) - \ln(0.0845)\} = 0.1378$.

Definition 1 is inspired by Liu and Chi (2001), but it remains sensible when there is no $\bar{\alpha}_h$ passing through (p_1, p_2), and also when there is more than one such $\bar{\alpha}_h$. Presupposing the availability of a two-stage test for each possible level, Liu and Chi proved that a two-stage level α test is equivalent to checking whether its overall p-value is not greater than α, and that by this property, the overall p-value is unique. In our terms, this is because the presupposed tests and the definition of the overall p-value are both based on the same choice of how to fill the unit square, i.e., the same choice of $\mathfrak{a} \in \mathfrak{A}$. This is formulated more precisely in the following lemma.

Lemma 1. *Let* $(\bar{\alpha}_h)_{h\in[0,1]} \in \mathfrak{A}$, $0 \le \alpha_1 \le \alpha_0 \le 1$, $\alpha \in [0,1]$ *and* $(p_1, p_2) \in [0,1]^2$.

(1) *If* $\alpha \notin [\alpha_1, \alpha_0]$, *or* $\alpha \in [\alpha_1, \alpha_0]$ *and* $p_1 \notin (\alpha_1, \alpha_0]$, *then* $p(p_1, p_2) \le \alpha$ *if and only if* $p_1 \le \alpha$.

(2) *If* $\alpha \in [\alpha_1, \alpha_0]$ *and* $p_1 \in (\alpha_1, \alpha_0]$, *then* $p(p_1, p_2) \le \alpha$ *if and only if* $p_2 \le \bar{\alpha}_{\alpha_2}(p_1)$, *where* α_2 *is arbitrary in* $[0,1]$ *with* $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(x)\,dx = \alpha$. *An* α_2 *satisfying this condition always exists, and for any two such* α_2 *and* α_2', $\bar{\alpha}_{\alpha_2} = \bar{\alpha}_{\alpha_2'}$ *on* $(\alpha_1, \alpha_0]$.

Therefore, assuming $\alpha \in [\alpha_1, \alpha_0]$, Method 1 is equivalent to rejecting the null hypothesis if and only if $p(p_1, p_2) \le \alpha$. The case $\alpha \notin [\alpha_1, \alpha_0]$ is included in Lemma 1 with regard to Proposition 2.

Proposition 2. *If* p_1 *and* p_2 *are independent and uniformly distributed on* $[0,1]$, *then* $p(p_1, p_2)$ *is uniformly distributed on* $[0,1]$. *If* p_1 *and* p_2 *are p-clud, then the distribution of* $p(p_1, p_2)$ *is stochastically not smaller than the uniform distribution on* $[0,1]$.

Proposition 2 has an important implication. The recursive combination principle of Brannath, Posch and Bauer (2002) is applicable, and multistage tests can be constructed.

The overall p-value function proposed by the same authors was based on the combination function C. Property 4 shows that it coincides with our notion in a special case.

Property 4. *Let* $\mathfrak{a} \in \mathfrak{A}$ *be induced by a p-value combination function* C *such that* $C(p_1, \cdot)$ *is left continuous for all* $p_1 \in (0,1)$. *The overall p-value can then be written as*

$$p(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \le \alpha_1 \text{ or } p_1 > \alpha_0, \\ \alpha_1 + \displaystyle\int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \le C(p_1,p_2)\}}\,dy\,dx & \text{otherwise,} \end{cases}$$

where $\mathbf{1}_{\{C(x,y)\le C(p_1,p_2)\}}$ *equals* 1 *if* $C(x,y) \le C(p_1,p_2)$, *and* 0 *otherwise.*

5. Example

We emphasize the idea of viewing a two-stage test as a family of conditional error functions that fills the unit square. According to Proposition 1, a (regular) p-value combination function C contains "too much" information: only the (generalized) level curves of the $C(p_1,p_2)$-surface are of interest. On the other hand, a single conditional error function $\bar{\alpha}$ is obviously "not enough". Only a family $\mathfrak{a} \in \mathfrak{A}$ provides the means to construct two-stage tests to any level, and to define overall p-values for two-stage tests.

Clearly, there are many ways to specify such a family. For instance, the most prominent conditional error functions, such as for Fisher's combination test or the inverse normal method, are already given as families. Alternatively, we may want to "extend" a single conditional error function $\bar{\alpha}$ to a family. This can be done in numerous ways, but it might be reasonable to pick an extension method that mimics the structure of a well-established test in some sense. As regards Fisher's combination test, this is particularly easy: define $\bar{\alpha}_r(x) = (\bar{\alpha}(x^r))^{1/r}$ for $r > 0$. When reparameterized by their integrals and completed by the constants $\bar{\alpha}_0 = 0$ and $\bar{\alpha}_1 = 1$, these functions form an element $\mathfrak{a}$ of $\mathfrak{A}$ (except if $\bar{\alpha} = 0$ or $\bar{\alpha} = 1$). In this context, however, it is more convenient to stick with the parameterization by $r > 0$. Indeed, Fisher's combination test is closed under this transformation. Starting with any conditional error function $\bar{\alpha}(x) = \min\{1, c_{\alpha_2}/x\}$ to a particular level α_2, the whole family is restored by $\bar{\alpha}_r(x) = \min\{1, c_{\alpha_2}^{1/r}/x\}$, $r > 0$.

If the initial function $\bar{\alpha}$ is not given, we are free to choose it as well. Following Adcock (1960), we may want the p-values of the two stages to cancel out in a symmetric fashion if early stopping is not considered: $p_2 = 1 - p_1$ should result in an overall p-value of 0.5. Thus, we apply the above transformation to the diagonal $y = 1 - x$. This yields the functions $\bar{\alpha}_r(x) = (1 - x^r)^{1/r}$, $r > 0$. A two-stage test based on this family is implemented according to Method 1, with the condition $p_2 \le \bar{\alpha}_{\alpha_2}(p_1)$ written as $p_1^{r(\alpha_2)} + p_2^{r(\alpha_2)} \le 1$, and $r(\alpha_2) > 0$ such that $\int (1 - x^{r(\alpha_2)})^{1/r(\alpha_2)}\, dx = \alpha_2$. Table 1 compares this new procedure to Fisher's combination test and the inverse normal method. It shows α_1 depending on α_0, and on the apportionment of α over the stages to satisfy the level condition in Method 1 for the level $\alpha = 0.05$. The case $\alpha_0 = 0.5$ corresponds to stopping for futility after the first stage if the observed effect shows in the wrong direction. The case $\alpha_0 = 1$ prohibits any stopping for futility. If $\alpha_2 = \alpha$, the full level is used after the final stage; if $\alpha_2 = \alpha_1$, the same local level is used after both stages (this is the "Pocock-type").

Table 1. Comparison of three two-stage tests. The table shows α_1 for $\alpha_0 = 0.5$ or $\alpha_0 = 1$, and for the full level after the final stage ($\alpha_2 = \alpha$) or the Pocock-type ($\alpha_2 = \alpha_1$), assuming $\alpha = 0.05$. FCT: Fisher's combination test. INM: inverse normal method. New: based on the family of conditional error functions $\bar{\alpha}_r(x) = (1 - x^r)^{1/r}$, $r > 0$.

		FCT	INM	New
$\alpha_2 = \alpha$	$\alpha_0 = 0.5$	0.0233	0.0044	0.0032
	$\alpha_0 = 1$	0.0087	0	0
$\alpha_2 = \alpha_1$	$\alpha_0 = 0.5$	0.0349	0.0307	0.0304
	$\alpha_0 = 1$	0.0323	0.0304	0.0302

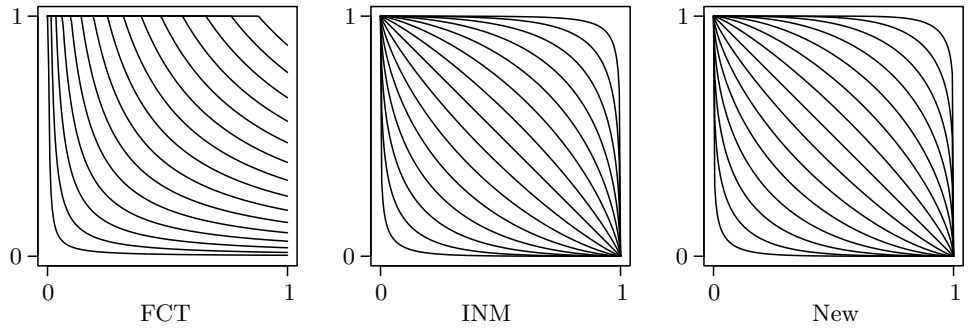

Figure 3. Three families of conditional error functions. FCT: Fisher's combination test. INM: inverse normal method. New: $\bar{\alpha}_r(x) = (1 - x^r)^{1/r}$, $r > 0$.

Note the similarity between the inverse normal method and the new test, especially for the Pocock-type. Indeed, the underlying families of conditional error functions look almost identical, as shown in Figure 3. If the full level α is used after the second stage and no stopping for futility is allowed ($\alpha_2 = \alpha$ and $\alpha_0 = 1$), these two tests never stop after the first stage. In the same situation, Fisher's combination test always rejects the null hypothesis when $p_1 \leq 0.0087$, and, strictly speaking, 0.0087 is just an upper bound (but of course a sensible choice) for α_1.

Tables 2 and 3 provide α_1 for the new test in a wider range of situations. The case $\alpha_1 = \alpha = \alpha_0 = 0.1$ is a single stage test. Particularly in the Pocock-type, α_0 matters only when small. This is because the area under the conditional error function becomes very small towards the right side of the unit square.

Table 2. Two-stage test based on the family of conditional error functions $\bar{\alpha}_r(x) = (1-x^r)^{1/r}$, $r > 0$, with the full level after the final stage. The table shows α_1 for different choices of α and α_0, under the condition $\alpha_2 = \alpha$.

α_0 \ α	0.1	0.05	0.025	0.01
			α_1	
0.1	0.1000	0.0365	0.0129	0.0031
0.2	0.0652	0.0205	0.0062	0.0012
0.3	0.0422	0.0115	0.0030	0.0005
0.4	0.0264	0.0063	0.0014	0.0002
0.5	0.0156	0.0032	0.0006	< 0.0001
0.6	0.0084	0.0015	0.0002	< 0.0001
0.7	0.0039	0.0006	< 0.0001	< 0.0001
0.8	0.0014	0.0002	< 0.0001	< 0.0001
0.9	0.0002	< 0.0001	< 0.0001	< 0.0001
1	0	0	0	0

Table 3. Two-stage test based on the family of conditional error functions $\bar{\alpha}_r(x) = (1-x^r)^{1/r}$, $r > 0$, with the same local level after both stages. The table shows α_1 for different choices of α and α_0, under the condition $\alpha_2 = \alpha_1$.

α_0 \ α	0.1	0.05	0.025	0.01
			α_1	
0.1	0.1000	0.0399	0.0174	0.0062
0.2	0.0767	0.0335	0.0154	0.0058
0.3	0.0690	0.0315	0.0149	0.0057
0.4	0.0657	0.0307	0.0147	0.0057
0.5	0.0642	0.0304	0.0146	0.0056
0.6	0.0635	0.0302	0.0146	0.0056
0.7	0.0632	0.0302	0.0146	0.0056
0.8	0.0631	0.0302	0.0146	0.0056
0.9	0.0631	0.0302	0.0146	0.0056
1	0.0631	0.0302	0.0146	0.0056

6. Discussion

Adaptive tests offer great flexibility in the planning and conduct of, for example, clinical trials. In recent times, procedures have been developed that do not even require a full prespecification of the test statistic or of the null hypothesis. While desirable in theory, such flexibility may be dangerous in practice. It does not open the door to total arbitrariness, but actually requires even more careful

study planning. The issue to be answered by the study should be thoroughly formulated. Still, when responsibly used, adaptive tests are a very versatile and practical tool.

The current article provides a formal link between two approaches to adaptive two-stage tests, namely, the *p-value combination function approach* and the *conditional error function approach*, in a general framework. The main idea is to view a two-stage test as a family of conditional error functions that fills the unit square. This family is used to define overall p-values in a way that covers previously given definitions based on either of the two approaches. In addition, new two-stage tests can be specified based on the same reasoning. The construction of multistage tests is possible by recursive combination as described by Brannath, Posch and Bauer (2002). These authors also outline the principles to construct point estimates and confidence intervals.

It is understood that the properties of a two-stage test and the meaningfulness of an overall p-value are highly dependent on the choice of the underlying family of conditional error functions. This family can have a vast variety of shapes. It remains to be explored which choices are advantageous from a practical perspective (see Brannath and Bauer (2004) for an investigation on "optimal" conditional error functions).

Acknowledgements

The author would like to thank two anonymous referees for their thorough review and helpful comments. He is also grateful to Rainer Schwabe (Otto-von-Guericke-Universität Magdeburg) for his very valuable suggestions.

Appendix A. Generalized Level Curves

The following points provide an insight into the relationship between a p-value combination function C and the corresponding conditional error functions $\bar{\alpha}^H$, $H \in \mathbb{R}$, as defined in Property 1.

(1) $C(p_1, p_2) \leq H$ implies $p_2 \leq \bar{\alpha}^H(p_1)$. The converse is true if $C(p_1, \cdot\,)$ is left continuous.

(2) $C(p_1, p_2) \geq H$ implies $p_2 \geq \bar{\alpha}^H(p_1)$ if $C(p_1, \cdot\,)$ is strictly increasing. The converse is true if $C(p_1, \cdot\,)$ is right continuous.

(3) If $C(p_1, \cdot\,)$ is continuous and strictly increasing, then

$$\bar{\alpha}^H(p_1) = \begin{cases} 0 & \text{if } C(p_1, p_2) > H \text{ for all } p_2 \in (0,1), \\ p_2 & \text{if there is a } p_2 \in (0,1) \text{ with } C(p_1, p_2) = H \\ & \text{(any } p_2 \text{ satisfying this condition is unique)}, \\ 1 & \text{if } C(p_1, p_2) < H \text{ for all } p_2 \in (0,1). \end{cases}$$

Appendix B. The Boundary of the Unit Square

Technical difficulties can arise for $p_1 \in \{0,1\}$ or $p_2 \in \{0,1\}$. For example, Method 1 does not cover the case $p_1 = \alpha_0 = 1$ since the $\bar{\alpha}_h$ are defined only on $(0,1)$. Similar problems appear in the context of overall p-values in Section 4. In many cases these difficulties can be avoided by defining C on $[0,1]^2$ or $\bar{\alpha}$ on $[0,1]$. We have settled for the smaller domains because this yields the link between the p-value combination function approach and the conditional error function approach in the most general form. It also covers such examples as the inverse normal method, where C tends to infinity towards the boundary of the unit square. In most applications this type of problem occurs only with probability 0.

Appendix C. Proofs

We write $x_n \uparrow x$ ($x_n \downarrow x$) if a sequence $(x_n)_n$ is nondecreasing (nonincreasing) and convergent with limit x.

C.1. Proof of Property 1

The function $\bar{\alpha}^H$ is nonincreasing since $C(\,\cdot\,, p_2)$ is nondecreasing for all p_2. To show that $\bar{\alpha}^H$ is left continuous, take a sequence $p_n \uparrow p_1$. Then $\bar{\alpha}^H(p_1) \leq \lim \bar{\alpha}^H(p_n)$. If $\bar{\alpha}^H(p_1) < \lim \bar{\alpha}^H(p_n)$, there would be some p_2 such that $\bar{\alpha}^H(p_1) < p_2 < \bar{\alpha}^H(p_n)$, and thus $C(p_1, p_2) > H$ and $C(p_n, p_2) \leq H$, for all n. This, however, cannot be since $C(\,\cdot\,, p_2)$ is left continuous.

If $H \leq H'$, then clearly $\bar{\alpha}^H \leq \bar{\alpha}^{H'}$. We finally show that $\int \bar{\alpha}^H(p_1)\, dp_1 \leq \int \bar{\alpha}^{H'}(p_1)\, dp_1$ implies $\bar{\alpha}^H \leq \bar{\alpha}^{H'}$; the converse is obvious. Take p such that $\bar{\alpha}^{H'}(p) < \bar{\alpha}^H(p)$, and let $\epsilon = (\bar{\alpha}^H(p) - \bar{\alpha}^{H'}(p))/2$. Necessarily $H' \leq H$ and $\bar{\alpha}^{H'} \leq \bar{\alpha}^H$. Since $\bar{\alpha}^{H'}$ is left continuous, there exists $p' < p$ such that $\bar{\alpha}^{H'}(p_1) < \bar{\alpha}^H(p) - \epsilon$ for all $p_1 \in [p', p]$. Thus, $\int_{p'}^{p} \bar{\alpha}^{H'}(p_1)\, dp_1 < \bar{\alpha}^H(p)(p - p') \leq \int_{p'}^{p} \bar{\alpha}^H(p_1)\, dp_1$, and therefore $\int \bar{\alpha}^{H'}(p_1)\, dp_1 < \int \bar{\alpha}^H(p_1)\, dp_1$.

The remark about existence of $\bar{\alpha}_0$ or $\bar{\alpha}_1$ is straightforward to prove.

C.2. Proof of Property 2

Let $C(p_1, p_2) = \inf\{h \in [0,1];\ \bar{\alpha}_h(p_1) \geq p_2\}$. It is easily seen that C cannot be infinite, and that it is nondecreasing in both arguments. To show that $C(\,\cdot\,, p_2)$ is left continuous, take a sequence $p_n \uparrow p_1$. Then $\lim C(p_n, p_2) \leq C(p_1, p_2)$. If $\lim C(p_n, p_2) < C(p_1, p_2)$, there would be some h such that $C(p_n, p_2) < h < C(p_1, p_2)$, and thus $\bar{\alpha}_h(p_n) \geq p_2$ and $\bar{\alpha}_h(p_1) < p_2$, for all n. But this cannot be since $\bar{\alpha}_h$ is left continuous.

Using the left continuity of the $\bar{\alpha}_h$, it can be shown that $\bar{\alpha}_h(p_1)$ is a right continuous function in h for fixed p_1. Therefore, $C(p_1, p_2) = \min\{h \in [0,1];\ \bar{\alpha}_h(p_1) \geq p_2\}$.

C.3. Proof of Property 3

For $\mathfrak{a} = (\bar{\alpha}_h)_h$ and $C = \tilde{C}(\mathfrak{a})$ as in Property 2, we show $\bar{\alpha}_h(p_1) = \max\{\sup[p_2 \in (0,1);\, C(p_1,p_2) \le h],\, 0\}$. Suppose $\bar{\alpha}_h(p_1) \in (0,1)$; the case $\bar{\alpha}_h(p_1) \in \{0,1\}$ can be treated in a similar way. Clearly $C(p_1, \bar{\alpha}_h(p_1)) \le h$, and therefore $\bar{\alpha}_h(p_1) \le \max\{\sup[p_2 \in (0,1);\, C(p_1,p_2) \le h],\, 0\}$. Assume $\bar{\alpha}_h(p_1) < \max\{\sup[p_2 \in (0,1);\, C(p_1,p_2) \le h],\, 0\}$. Then there would exist $p_2 > \bar{\alpha}_h(p_1)$ such that $C(p_1,p_2) \le h$. This, however, would yield $\bar{\alpha}_h(p_1) \ge \bar{\alpha}_{C(p_1,p_2)}(p_1) \ge p_2$.

C.4. Proof of Proposition 1

Proposition 1 follows directly from Properties 1–3.

C.5. Proof of Lemma 1

Only the case $\alpha \notin [\alpha_1, \alpha_0]$ and $p_1 \in (\alpha_1, \alpha_0]$ needs to be considered in (1). If $\alpha < \alpha_1$, then both $p(p_1,p_2)$ and p_1 are greater than α. If $\alpha > \alpha_0$, then both $p(p_1,p_2)$ and p_1 are smaller than α.

To prove the existence of α_2 in (2), let $\gamma = \inf(A)$ and $\gamma' = \sup(B)$ for $A = \{h \in [0,1];\, \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_h(x)\,dx \ge \alpha\}$, $B = \{h \in [0,1];\, \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_h(x)\,dx \le \alpha\}$. It is not difficult to show $\gamma \in A$ and $\gamma' \in B$. Clearly, $\gamma \le \gamma'$. If $\gamma < \gamma'$, then $\alpha - \alpha_1 \le \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_\gamma(x)\,dx \le \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\gamma'}(x)\,dx \le \alpha - \alpha_1$, so γ and γ' both satisfy the condition required for α_2. The case $\gamma = \gamma'$ is obvious. The uniqueness of $\bar{\alpha}_{\alpha_2}$ on $(\alpha_1, \alpha_0]$ can be shown by arguments similar to those at the end of the proof of Property 1.

Now assume $p_2 \le \bar{\alpha}_{\alpha_2}(p_1)$ with α_2 such that $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(x)\,dx = \alpha$, and let $h^\star = \min\{h \in [0,1];\, \bar{\alpha}_h(p_1) \ge p_2\}$ as in Definition 1. Then obviously $h^\star \le \alpha_2$, and thus $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{h^\star}(x)\,dx \le \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(x)\,dx = \alpha$. We omit the details for the converse.

C.6. Proof of Proposition 2

Suppose p_1 and p_2 are independent and uniformly distributed on $[0,1]$. If $\alpha \notin [\alpha_1, \alpha_0]$, then $\Pr(p(p_1,p_2) \le \alpha) = \alpha$ by Lemma 1(1). If $\alpha \in [\alpha_1, \alpha_0]$, there is α_2 such that $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(x)\,dx = \alpha$. By Lemma 1(1) and (2), $\Pr(p(p_1,p_2) \le \alpha) = \Pr(p_1 \le \alpha_1) + \Pr(\{p_1 \in (\alpha_1,\alpha_0)\} \cap \{p_2 \le \bar{\alpha}_{\alpha_2}(p_1)\}) = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{\alpha_2}(x)\,dx = \alpha$. If p_1 and p_2 are p-clud, $\Pr(p(p_1,p_2) \le \alpha) \le \alpha$ by a similar argument.

C.7. Proof of Property 4

Let $h^\star = \min\{h \in [0,1];\, \bar{\alpha}_h(p_1) \ge p_2\}$ and $H = C(p_1,p_2)$. By Appendix

A(1) it can be shown that $h^\star = \int \bar{\alpha}^H(x)\,dx$, and for $p_1 \in (\alpha_1, \alpha_0]$,

$$\begin{aligned} p(p_1, p_2) &= \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}_{h^\star}(x)\,dx \\ &= \alpha_1 + \int_{\alpha_1}^{\alpha_0} \bar{\alpha}^H(x)\,dx \\ &= \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{y \le \bar{\alpha}^H(x)\}}\,dy\,dx \\ &= \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \le H\}}\,dy\,dx. \end{aligned}$$

C.8. Proof of Appendix A.

To show (1), note that $C(p_1, p_2) \le H$ implies $p_2 \le \bar{\alpha}^H(p_1)$ by the definition of $\bar{\alpha}^H$. Now let $C(p_1, \cdot\,)$ be left continuous, and suppose $p_2 \le \bar{\alpha}^H(p_1)$. If $p_2 < \bar{\alpha}^H(p_1)$, there is some $p_2' > p_2$ with $C(p_1, p_2') \le H$, so $C(p_1, p_2) \le C(p_1, p_2') \le H$. If $p_2 = \bar{\alpha}^H(p_1)$, take a sequence $p_n \uparrow p_2$ with $C(p_1, p_n) \le H$ for all n. Since $C(p_1, \cdot\,)$ is left continuous, $C(p_1, p_2) \le H$. (2) can be shown similarly. If $C(p_1, \cdot\,)$ is continuous and strictly increasing, then $p_2 = \bar{\alpha}^H(p_1) \Leftrightarrow C(p_1, p_2) = H$ due to (1) and (2). This proves (3).

References

Adcock, C. J. (1960). A note on combining probabilities. *Psychometrika* **25,** 303-305.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132,** 235-244.

Bauer, P. (1989a). Multistage testing with adaptive designs (with discussion). *Biom. und Inform. in Med. und Biol.* **20,** 130-148.

Bauer, P. (1989b). Sequential tests of hypotheses in consecutive trials. *Biometrical J.* **31,** 663-676.

Bauer, P., Brannath, W. and Posch, M. (2001). Flexible two-stage designs: an overview. *Methods Inf. Med.* **40,** 117-121.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029-1041. Correction in *Biometrics* **52** (1996), 380.

Bauer, P. and Röhmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statist. Medicine* **14,** 1595-1607.

Brannath, W. and Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. *Biometrics* **60,** 715-723.

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *J. Amer. Statist. Assoc.* **97**, 236-244.

Chi, G. Y. H. and Liu, Q. (1999). The attractiveness of the concept of a prospectively designed two-stage clinical trial. *J. Biopharm. Statist.* **9**, 537-547.

Cui, L., Hung, H. M. J. and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55,** 853-857.

DeMets, D. L. and Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* **67,** 651-660.

DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69,** 661-663.

Fairbanks, K. and Madsen, R. (1982). *P* values for tests using a repeated significance test design. *Biometrika* **69,** 69-74.

Fisher, L. D. (1998). Self-designing clinical trials. *Statist. Medicine* **17,** 1551-1562.

Fisher, R. A. (1932). *Statistical Methods for Research Workers.* Oliver & Boyd, London.

Gould, A. L. and Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69,** 75-80.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statist. Medicine* **22,** 971-993.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

Lan, K. K. G., Simon, R. and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Comm. Statist.–Sequential Analysis* **1,** 207-219.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55,** 1286-1290.

Li, G., Shih, W. J., Xie, T. and Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3,** 277-287.

Liu, Q. and Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57,** 172-177.

Mosteller, F. and Bush, R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology* (Edited by G. Lindzey), 289-334. Addison-Wesley, Cambridge.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57,** 886-891.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Plann. Inference* **42**, 19-35.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191-199.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical J.* **41,** 689-696.

Proschan, M. A. (2003). The geometry of two-stage tests. *Statist. Sinica* **13,** 163-177.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315-1324.

Shen, Y. and Fisher, L. D. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55,** 190-197.

Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40,** 797-803.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Statist.* **16,** 117-186.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19,** 326-339.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43,** 193-199.

Wassmer, G. (1999). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien.* Verlag Alexander Mönch, Köln.

Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statist. Papers* **41**, 253-279.

Schering AG, D-13342 Berlin, Germany.

E-mail: vandemem@gmx.de

(Received September 2004; accepted April 2005)