Evaluating the Operating Characteristics of a Class of Closed Multiple Comparisons Procedures: A General Framework

December 15, 2001
Running head: Power of Closed Tests

Keywords: high dimensional integration, quasi Monte Carlo, multiple testing, directional errors, familywise error, predictive power
Abstract

Analytic formulas are developed for various types of power and error rates of some closed testing procedures. The formulas involve non-convex regions that may be integrated with high, pre-specified accuracy using available software. The non-convex regions are represented as a union of hyper-rectangles. These regions are transformed to the unit hypercube, then summed, to create an expression for power that is the integral of a function defined on the unit hypercube. This function is then evaluated using existing quasi-Monte Carlo methods that are known for their accuracy. Applications include individual, average, complete, and minimal power, for closed pairwise comparisons (max T-based), as well as fixed sequence and non-pairwise comparisons. An extension to Bayesian predictive power also is given. The methodology extends naturally to the evaluation of combined directional and non-directional Type I error rates, which we investigate thoroughly and find no evidence of excesses in applications involving the noncentral multivariate t distribution.

1 Introduction and Notation

The closed testing procedures (CTPs) of Marcus, Peritz and Gabriel (1976) cover a variety of useful methods, including stepwise and fixed sequence tests. These methods are known to have good power (Grechanovsky and Hochberg, 1999), and can be specifically tailored to applications ranging from clinical trials to genetics (see, for example, Bauer, Röhmel, Maurer and Hothorn 1998; and Westfall, Zaykin and Young, 2001, respectively).
Our main concern is in testing multiple null hypotheses $H_j : \theta_j = 0$ versus their two-sided alternatives $K_j : \theta_j \neq 0$, $j = 1, \ldots, k$. A CTP requires pre-specified tests for all intersection nulls of the form $H_S = \cap_{j \in S} H_j$, where $S$ is a non-empty subset of $\{1, \ldots, k\}$. Let $S$ denote the set of all intersections (removing redundancies that can occur, e.g., in pairwise comparisons), and specify an $\alpha$-level test for each $H_S$, where $S \in S$. Then, using the CTP, one rejects any $H_j$ only if all $H_S$ are rejected, for $S \supseteq \{j\}$.

Thus, CTPs are relatively simple to implement for data analysis: one needs only to specify $\alpha$-level tests for the various possible intersection hypotheses, then proceed algorithmically as described above. However, it can be much more difficult to evaluate their operating characteristics such as power and level, due to their complicated critical regions.

Power functions for such procedures can be evaluated in some special cases, notably in cases involving pairwise comparisons against a control (Dunnett, Horn and Vollandt, 2001) or in cases of single-step testing procedures for which the power functions are simplified considerably (Horn, Vollandt and Dunnett, 2000). However, in more general cases, including the important case of all pairwise comparisons, analytic methods for evaluating power are heretofore unavailable.

Further, while familywise Type I error rate (FWE) control is established quite simply for CTPs, there are pathological cases where the directional error rate and type I error rate (called “combined error rate,” or CER, below) is uncontrolled (Shaffer, 1980; Liu, 1997). Finner (1999) has demonstrated CER control for certain cases of practical interest, but it is unknown whether such results hold more generally. Specifically, suppose that some (possibly
empty) subset of nulls is true, and define $A_1$ to be the event that at least one of the true nulls is rejected (i.e., that there is at least one type I error). Then $\text{FWE} = P(A_1)$. Having rejected $H_j$, one naturally wishes to claim that the sign of $\theta_j$ is the same as that of its estimate $\hat{\theta}_j$. To make this claim, one requires control of both type I errors and errors in determining the sign of non-null effects. A sign error (also called type III error or directional error in the literature) is defined as a rejection of a false null, but where the sign of $\theta_j$ is opposite that of $\hat{\theta}_j$. Let $A_2$ be the event that there is at least one sign error among the true non-null effects, and define combined error rate (CER) as $\text{CER} = P(A_1 \cup A_2)$.

In this article we consider CTPs that result when $H_S$ is tested using “maxT” statistics usually of the form $\max_{j \in S} |T_j|$. Our general methodology allows extensions to tests of the form $\max_{j \in S'} |T_j|$, for $S' \subset S$, which allows directed tests of the form considered by Rom et al. (1994); it also allows maxT tests using one-sided statistics. MaxT-type tests are popular for several reasons. First, they are closely related to Tukey’s and Dunnett’s method for pairwise comparisons, both of which are based on the distribution of the maxT statistic. Second, use of the maxT test with Bonferroni critical values leads to the popular stepwise method pioneered by Holm (1979) for free combinations (e.g., Dunnett-type comparisons), which was later extended by Shaffer (1986) to restricted combinations (e.g., all pairwise comparisons). Westfall (1997) extended Shaffer’s method to incorporate correlations between tests, providing more power. While not uniformly most powerful, maxT-based joint tests have good power in the case where many of the tests are near null and one is markedly non-null; such alternatives might be expected in screening studies, for example.
The distributional set-up is as follows. Suppose the $\theta_j$ are estimated using $\hat{\theta}_j$, let $\theta' = (\theta_1, \ldots, \theta_k)$, and suppose $\hat{\theta} \sim N_k(\theta, \sigma^2 V)$, where $V$ is a known positive semi-definite matrix with positive diagonals. Further, suppose an independent estimate $\hat{\sigma}^2$ is available such that $\nu \hat{\sigma}^2 / \sigma^2 \sim \chi^2_\nu$, and define $T_j = \hat{\theta}_j / (v_{jj} \hat{\sigma}^2)^{1/2}$, where $v_{jj}$ is the $j$th diagonal of $V$. Then $T' = (T_1, \ldots, T_k)$ has the multivariate t distribution with parameters $\delta$, $R$, and $\nu$, where the noncentrality vector $\delta$ has elements $\delta_j = \theta_j / (v_{jj} \sigma^2)^{1/2}$, the dispersion matrix is $R = (\text{diag}\{V\})^{-1/2}V(\text{diag}\{V\})^{-1/2}$, and where the degrees of freedom are $\nu$. These assumptions cover a range of applications of interest, including all pairwise comparisons, comparisons with a control, general contrasts, fixed-sequence, and multiple one-sided tests. In this article we evaluate power and error rates for such CTPs with high accuracy, facilitating (a) comparisons of CTPs with other methods, (b) sample size selection when CTPs are planned, and (c) evaluation of CERs.

Section 2 introduces two different methods of deriving the appropriate integration regions particular to a given application. Numerical issues concerning the computation of sums of multivariate noncentral t probabilities with arbitrary correlation matrices are covered in Section 3. The following Sections 4 and 5 expand the scope of applications to general CTPs including all-pairwise comparisons, fixed sequence tests and predictive power calculation. In addition, ways of evaluating CERs precisely for a variety of configurations are developed. Section 6 summarizes the results and sketches some further fields of applications.
2 Representation of the Integration Regions

2.1 Hyper-Rectangles of CTPs that use maxT Tests

While the rejection regions associated with CTPs are complex, they can always be decomposed into finite unions of disjoint hyper-rectangles when the maxT tests are used. To see this, note that there are no more than $|S|$ distinct critical values $c_S$ for the maxT-based CTP, where $H_S$ is rejected when $\max_{j \in S} |T_j| \geq c_S$ (dependence of the critical values on $\alpha$ is suppressed for notational convenience). Let $m$ denote the number of distinct critical values, labeled $c_1 < \cdots < c_m$, and define $c_0 = 0$ and $c_{m+1} = \infty$. Then the $k$-dimensional positive orthant may be partitioned (excepting a set of measure 0) into $(m + 1)^k$ hyper-rectangles of the form $C_i = \prod_{j=1}^k (c_{i_j}, c_{i_j} + 1)$, where $i_j \in \{0, 1, \ldots, m\}$ and $i = (i_1, \ldots, i_k)$. Noting that $\max_{j \in S} c_{i_j} < \max_{j \in S} t_j < \max_{j \in S} c_{i_j} + 1$, for all $t = (t_1, \ldots, t_k) \in C_i$, we see that $H_S$ will be rejected for all $t \in C_i$ if $\max_{j \in S} c_{i_j} \geq c_S$; similarly, $H_S$ will be accepted for all $t \in C_i$ if $\max_{j \in S} c_{i_j} + 1 \leq c_S$. Thus, the decision to reject or accept $H_j$, when using the CTP, is constant over the hyper-rectangle $C_i$.

The regions $C_i$ concern absolute values of the $t$-statistics, and can be converted easily to hyper-rectangles for the actual $t$-statistics as described in the previous section, by expressing each $C_i$ as a union of at most $2^k$ disjoint hyper-rectangles. Specifically, if $c_{i_j} > 0$, then the coordinate range $(c_{i_j}, c_{i_j} + 1)$ can be expanded as $(-c_{i_j} + 1, -c_{i_j}) \cup (c_{i_j}, c_{i_j} + 1)$ for the actual $t$-statistics; if $c_{i_j} = 0$, then the coordinate range $(c_{i_j}, c_{i_j} + 1)$ can be expanded as $(-c_{i_j} + 1, c_{i_j} + 1)$. If there are $q$ zeros, then there are at most $2^{k-q}$ disjoint hyper-rectangles. Often, many
hyper-rectangles collapse, leaving fewer hyper-rectangles, as we will show in the examples.

Let $D_h^1, h = 1, \ldots, M_1$ denote a set of disjoint hyper-rectangles, for which a particular $H_j$ is rejected using a CTP. We then may represent $P(\text{Reject } H_j)$ as $\sum_{h=1}^{M_1} P(T \in D_h^1)$, which is a sum of integrals. The main idea is that this sum of integrals can simplified to a single integral over a unit hypercube as shown in Section 3. Using known methods for evaluating functions over the unit hypercube, we can thereby evaluate power very accurately.

### 2.2 Union-Intersection Representations

Alternatively, one can represent the CTP rejection regions using union-intersection formulas, often resulting in fewer summands. Let $\psi_S = \psi_S(T)$ denote the critical function for testing $H_S$, where $S \in \mathcal{S}$; that is, $\psi_S = 1$ if $\max_{T \in S} |T_i| \geq c_S$, otherwise $\psi_S = 0$. Then the corresponding critical function for testing $H_j$ using the CTP is $\Psi_j = \Pi_{S \supseteq \{j\}} \phi_S$, and the power function is given by $E(\Psi_j)$. We then represent $\Psi_j$ as $\Pi_{S \supseteq \{j\}}(1 - \tau_S)$, where $\tau_S$ is the indicator of $\max_{j \in S} |T_j| < c_S$. The resulting expansion of $\Psi_j$ in terms of the $\tau_S$ is a sum of $M_2$ elements with alternating signs and products of the form $\tau_{S_1} \tau_{S_2} \ldots$, whose expectations are integrals over regions $D_h^2$, $h = 1, \ldots, M_2$ (similar to those as described in Section 2.1). Here, $M_2$ is bounded by $\sum_{i=1}^{m'} \binom{1}{m} = 2^{m'} - 1$, where $m'$ is the number of sets with $S \supseteq \{j\}$ (again, regions often may be collapsed, as shown in the applications). Similar inclusion-exclusion inequalities have been used by Naiman and Wynn (1997) to obtain probability bounds for multiple comparisons problems; our focus is on finding exact probabilities.
3 Numerical Evaluation of Sums of Multivariate $t$ Probabilities

The rectangular regions $D_h^j, j = 1, 2$ described in Section 2 involve the multivariate $t$ distribution, which can be written as

$$2^{1 - \frac{\nu}{2}} \frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2} + \frac{d}{2}\right)} \int_0^\infty s^{\nu - 1} e^{-\frac{s^2}{2}} \Phi_d \left( \frac{sa - \delta}{\sqrt{\nu}}, \frac{sb - \delta}{\sqrt{\nu}}, R \right) ds,$$

where $\Phi_d(a, b, R) = \frac{1}{\sqrt{|R| (2\pi)^d}} \int_{a_1}^{b_1} \ldots \int_{a_d}^{b_d} e^{-\frac{1}{2}x^T R^{-1} x} dx$ denotes the $d$-variate normal distribution, $x = (x_1, x_2, \ldots, x_d)^T$ and $-\infty \leq a_i < b_i \leq \infty$ for $i = 1, \ldots, d$. The integration region $[a, b] = [a_1, b_1] \times \ldots \times [a_d, b_d]$ is given through the particular hyper-rectangle $D_h^1$ or $D_h^2$ to be integrated. The total number $M$ of integral expressions (either $M_1$ or $M_2$) depends on the selected representation.

Genz and Bretz (2001) described quasi-Monte Carlo methods for efficiently integrating the noncentral $t$ distribution. Each integral of the form (1) is transformed to one over the unit hypercube

$$T_d(a, b) = \int_0^1 \int_0^1 \ldots \int_0^1 g(w) dw,$$

where $g(w) = \prod_{i=1}^d (e_i - d_i)$ defines the new integrand. Let $R = \tilde{C} \tilde{C}'$ be the Cholesky decomposition of $R$ and denote by $\chi_{\nu}^{-1}$ the inverse $\sqrt{\chi^2/\nu}$-distribution function, with $\tilde{C}$ being a lower triangular matrix. Then, $e_{i+1} = \Phi_1 \left( \chi_{\nu}^{-1} (w_d) \left( b_{i+1} - \sum_{j=1}^i \tilde{c}_{i+1,j} y_j \right) / (\tilde{c}_{i+1,i+1} \sqrt{\nu}) \right)$ for $i = 1, \ldots, d - 1$, where $e_1 = \Phi_1 \left( b_1 \chi_{\nu}^{-1} (w_d) (\tilde{c}_{11} \sqrt{\nu}) \right), y_i = \Phi^{-1} \left( d_i + w_i (e_i - d_i) \right)$, and $\Phi_1$ is the cumulative distribution function.
\textbf{w} = (w_1, \ldots, w_d)^t \in [0,1]^d$. The $d_i$’s are defined correspondingly for the lower bound $a$. Genz and Bretz (2001) show that the application of randomized lattice rules leads to reliable evaluations of the integrals within three or four significant digits, all in reasonable computation time. These methods perform particularly good in comparison to standard Monte Carlo integration methods. The total number of function evaluations is substantially reduced so that the repeated, error-prone numerical computation of $\Phi$, $\Phi^{-1}$ and $\chi^{-1}$ does not impact the overall accuracy of the proposed integration method. For a more detailed description of lattice rules and related techniques see Sloan and Joe (1994).

Due to possibly high values of $M$, it is inefficient to compute each integral individually, then sum the results, since the total integration error could only be bounded roughly by $M\varepsilon$, where $\varepsilon$ is the individual error tolerance per integral. Instead we consider

$$
\sum_{h=1}^{M} \int_{D_h} g_h(w) dw = \sum_{h=1}^{M} \int_{[0,1]^d} \prod_{i=1}^{d} (e_{hi}(w) - d_{hi}(w)) dw = \int_{[0,1]^d} \sum_{h=1}^{M} \prod_{i=1}^{d} (e_{hi}(w) - d_{hi}(w)) dw.
$$

(2)

Note that without loss of generalization all those integrals with dimension $d' < d$ are treated as $d$–dimensional integrals with the additional $d' - d$ components being integrated from $-\infty$ to $\infty$. Thus, the simultaneous evaluation of $M$ multivariate $t$ integrals can be reduced essentially to a single integral of dimension $d$. In particular, the resulting error bound is precise and needs no further adjustment for the multiplicity of integrals considered.

The algorithm allows singular cases where $|R| = 0$. Such situations occur frequently; for example, the all-pairwise comparisons case considered in the next section results in singular
R since the number of contrasts exceeds the dimensionality of the parameters. Genz and Kwong (2000) have shown for the multivariate normal case that in typical applications the effective dimension reduces to at most \( g - 1 \). Their method extends in a natural way to the present multivariate \( t \) problems.

### 4 Applications

#### 4.1 Power calculations for all pairwise comparisons

We investigate the application of equation (2) on different types of power: individual, average, minimal and complete power. Consider the all-pairwise comparisons \( \theta_j = \mu_i - \mu_{i'} \), \( 1 \leq i < i' \leq g \), \( j = 1, \ldots, k = g(g - 1)/2 \). For the purpose of illustration we restrict the representations on \( g = 3 \) here, but the techniques generalize to higher \( g \). The actual number \( M \) of integrals coming from either of the representations of Section 2 can be reduced as follows. First, conditional on \( \delta \), the values of two or more summands might be equal. Integrals of the form \( \Phi_d \left( \frac{a}{\sqrt{\nu}} - \delta, \frac{b}{\sqrt{\nu}} - \delta, R \right) \) yield the same values for different \([a, b]\), as long as the rectangles are obtained by permutation within each group of equal noncentrality parameters. Thus, several integrals can be summarized by one single computation. Second, the union of events \( E_{i_1}, \ldots, E_{i_p} \) may lead to the same probability expression as the union of a second set of events \( E_{i_1}, \ldots, E_{i_q} \). Relationships of the latter type usually lead to the cancel out of expressions of opposite signs. An algorithmic implementation of such refinements avoids the inclusion of a number of unnecessary expressions for the explicit power computation (2).
Individual Power

The individual power is defined as the probability of correctly rejecting a particular $H_j$, i.e. $\Pi_{ind}^j = P(\Phi_j = 1|K_j)$. Thus,

$$
\Pi_{ind}^j = P(|T_{123}| > c_2 \land |T_j| > c_1)
= 1 - P(|T_{123}| \leq c_2) - P(|T_j| \leq c_1) + P(|T_{123}| \leq c_2 \land |T_j| \leq c_1)
= 1 - \int \int \int f(x)dx - \int \int \int f(x)dx + \int \int \int f(x)dx
$$

where $T_{i_1,...,i_l} = \max\{T_{i_1},...,T_{i_l}\}$ for some subset $\{i_1,...,i_l\} \subseteq \{1,...,k\}$. Each integral without a specified integration region is defined over $\mathbb{R}$. The noncentral multivariate t pdf is denoted by $f$. For the sake of a simplified notation the assignment of the rectangles to the variables $x_j, j = 1,...,3$, is left slightly imprecise, and integrals over similar rectangles are grouped together. The integration region consists of a cube of half width $c_2$ and a strip of half width $c_1$ (see Figure 1a). The decomposition of the integration region using hyper-rectangles is not unique. Partitioning the region into disjoint subsets as described in Section 2.1 also leads to the sum of three integrals:

$$
\Pi_{ind}^j = 1 - \int \int \int f(x)dx - \int \int \int f(x)dx - \int \int \int f(x)dx.
$$

At this time, theory is lacking to determine the smallest set of disjoint rectangles for general $g$; however, one may obtain regions through the methods of sections 3.1 and 3.2 to identify and collapse regions, then choose the simplest representation. The authors have developed specialized software for this purpose in cases where $g > 3$. 

12
Average Power

Closely related to the individual power is the concept of average power. This is defined as the average expected number of correct rejections of all non-zero elementary hypotheses $K_{j_1}, \ldots, K_{j_i}$, $\Pi_{j_1, \ldots, j_i}^{\text{ave}} = \frac{1}{i} \sum_{l=1}^{i} \Pi_{j_l}^{\text{ind}}$. Thus, the computation reduces to a weighted sum of above integral expressions.

Minimal Power

The minimal power is defined as

$$\Pi_{j_1, \ldots, j_i}^{\text{min}} = P \left( \exists l = 1, \ldots, i : \Phi_{j_l} = 1 | \cap_{l=1}^{i} K_{j_l} \right).$$

If $i = 1$ (of possible relevance for other than all-pairwise comparisons), the minimal power reduces to the expression for $\Pi_{j_1}^{\text{ind}}$. For $i = k$, $\Pi_{j_1, \ldots, j_i}^{\text{min}} = P (|T_{1, \ldots, k}| > c_m)$. In all other cases the integration regions are more complex. For example, a direct application of the union-intersection representation for the case $g = 3$ and $i = 2$ leads to

$$\Pi_{j_1, j_2}^{\text{min}} = P \left( |T_{123}| > c_2 \wedge \left( |T_{j_1}| > c_1 \lor T_{j_2} > c_1 \right) \right)$$

$$= 1 - \int \int \int f(x)dx - \int \int \int f(x)dx + \int \int f(x)dx.$$

Due to the common variance estimate the same critical value is used for both $T_{j_1}$ and $T_{j_2}$.

The corresponding integration region, which consists of a single cube of half width $c_2$ and an intersecting tube of half width $c_1$, is depicted in Figure 1b. The partition of the region into disjoint hyper-rectangles leads also to the sum of three integrals.
Complete Power

Finally, integral expressions for the complete power, defined as

$$\Pi_{j_1, \ldots, j_i}^{\text{compl}} = P (\forall l = 1, \ldots, i : \Phi_{j_l} = 1 | K_{jl}) ,$$

are derived similarly using either of the methods proposed in Section 2. For $g = 3$ and $i = 2$ (necessarily $\delta = (\delta, \delta, 0)$ or any permutation of it), the union-intersection principle leads to an initial sum of 7 integrals. Symmetry conditions in the integration region, as given at the beginning of this section, reduce the initial upper bound of 7 integrals to the sum of 5 integrals,

$$\Pi_{j_1, j_2}^{\text{compl}} = P (|T_{123}| > c_2 \land |T_{j_1}| > c_1 \land |T_{j_2}| > c_1)$$

$$= 1 - \int \int \int f(x)dx - 2 \int \int f(x)dx + 2 \int \int \int f(x)dx$$

$$+ \int \int \int f(x)dx - \int \int \int f(x)dx .$$

Figure 1c shows a graphical representation of the acceptance region. If $i = 3$, relationships in the inclusion-exclusion formulas, such as $P (E) = P (E \cup \{|T_{123}| > c_2\})$ (where $E$ denotes the event $\bigcup_{l < l'}\{|T_{j_l, j_{l'}}| > c_1\}$), reduce the initial number of 15 integrals to 13 integrals by cancelling out expressions of opposite signs. Further reductions are possible. For example, in the case of two equal noncentrality parameters $\delta = (\delta_1, \delta_1, \delta_2), \delta_1 \neq \delta_2$, the above symmetry
considerations lead to a sum of nine expressions,

\[
1 - \Pi_{123}^{\text{compl}} = \int \int \int f(x)dx \quad + 2 \int \int \int f(x)dx \quad + \int \int \int f(x)dx \\
-2 \int \int \int f(x)dx \quad - \int \int \int f(x)dx \quad - \int \int \int f(x)dx \\
-2 \int \int \int f(x)dx \quad + \int \int \int f(x)dx \quad + 2 \int \int \int f(x)dx.
\]

include Figure 1 about here

Figure 1: Graphical representation of integration regions of different applications \((g = 3, \text{ details given in the text)}\): (a) individual power, (b) minimal power, (c) complete power, (d) combined error rates.

**Power gain of CTP over single-step methods**

A criticism of CTPs is that confidence intervals are not available to correspond with the test-based decisions. Tukey’s method for confidence intervals yields a single-step method that loses power relative to the CTP. As an application we computed the power differences between the single-step all-pairwise comparisons \((g = 4)\) and the corresponding CTP. Figure 2 gives the curves for the individual power of both procedures at two different sample sizes \((n = n_i = 10 \text{ and } n = n_i = 30 \forall i)\). Two different shifts expectation profiles were investigated. The first profile was of the form \((\Delta, 0, 0, 0)\), with values for the shift parameter \(\Delta\) ranging
from 0.1 to 1. The second profile was of the form \((3\Delta, 2\Delta, \Delta, 0)\). The plotted curves show the individual power for the comparison of the groups 1 and 2. As expected, the stepwise procedure is uniformly better, achieving up to 10% power gain in the former case and up to 20% in the latter case. Since the single step procedure is restricted to the single comparison of interest, the power curves are identical for both expectation profiles. The CTP is particularly more powerful for the second profile, since decisions regarding the other elementary hypotheses also enter the computations. The underlying integral expressions were derived in the same fashion as illustrated before for the case \(g = 3\).

*include Figure 2 about here*

**Figure 2:** Comparison of the individual power values for the single-step all-pairwise comparisons and the corresponding CTP; \(g = 4\), \(\alpha = 0.05\), \(\sigma = 1\), \(n_i = 10, 30\) with expectation profiles (a) \((\Delta, 0, 0, 0)\) and (b) \((3\Delta, 2\Delta, \Delta, 0)\).

**Power when hypotheses are rejected only “in the right direction”**

A different way of looking at above power considerations is to think about the power of two-sided tests being most sensibly calculated using only the side of interest. The idea is to exclude the cases where the hypotheses are rejected, but in the wrong direction, adding erroneously to more “power.” The numerical results of the foregoing sections still apply, although the power values remain practically unchanged. Let, for example, \(g = 3\) and \(K_{j_1}\) and \(K_{j_2}\) be the true alternatives with \(\delta_{j_1}, \delta_{j_2} > 0\). Correct directional decisions can be either
considered at all intermediate steps or they can be considered only at the level of testing the elementary hypotheses. For the calculation of the minimal power, the former case leads to

\[ P \left( (T_{j_1} > c_2 \lor T_{j_2} > c_2) \land (T_{j_1} > c_1 \lor T_{j_2} > c_1) \right) = P \left( T_{j_1} > c_2 \lor T_{j_2} > c_2 \right), \]

while the second case leads to

\[ P \left( |T_{123}| > c_2 \land (T_{j_1} > c_1 \lor T_{j_2} > c_1) \right) = 1 - \int \int \int_{[-c_2, c_2]^3} f(x) dx - \int \int f(x) dx \]

\[ + \int \int f(x) dx. \]

But as matter of fact, a few example calculations show that the differences between the various power definitions are indeed negligible. Let for example in the balanced case \( \delta = (2.2361, 2.2361, 0), \nu = 27 \) and \( \alpha = 0.05 \). Table 1 shows the differences between the various power definitions. The given power values are correct to four significant digits.

<table>
<thead>
<tr>
<th>Power definition</th>
<th>Probability expression</th>
<th>Minimal Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided decisions</td>
<td>( P \left(</td>
<td>T_{123}</td>
</tr>
<tr>
<td>Correct side decisions</td>
<td>( P \left( T_1 &gt; c_2 \lor T_2 &gt; c_2 \right) )</td>
<td>0.5711</td>
</tr>
<tr>
<td>Correct elementary decisions</td>
<td>( P \left(</td>
<td>T_{123}</td>
</tr>
</tbody>
</table>

Table 1: Power values for different definitions of power

### 4.2 Combined error rates

As discussed in the Introduction, the evaluation of CERs plays an important role in stagewise testing problems. Westfall, Tobias and Bretz (2000) investigated a variety of situations involving noncentral \( t \) distributions by means of Monte Carlo variance reduction methods. In the following we show how the integration results of the foregoing sections are applied to
the analytical evaluation of CERs.

For example, consider the successive comparisons $\theta_j = \mu_{j+1} - \mu_j, j = 1, \ldots, k = g - 1$ (Budde and Bauer, 1989). Assume that a single alternative $K_j$ is true. Let $E_1$ denote the event $|T_{1,2,\ldots,k}| < c_m$, $E_2 : T_j > c_m$ and $E_3 : \max_{i \not= j} |T_i| < c_{m-1}$. Thus we have \{no CER\} = \{E_1 \cup (E_2 \cap E_3)\}. Since $E_1$ and $E_2 \cap E_3$ are disjoint, we obtain with the notation of the Introduction,

$$P \left( A_1^c \cap A_2^c \right) = \int \cdots \int f(x) dx + \int \cdots \int f(x) dx.$$

For any $g$, the calculation involves the integration over the central cube of half width $c_m$ and an infinite tube of half width $c_{m-1}$ starting from the cube (see Figure 1d). If $P \left( A_1^c \cap A_2^c \right) \geq 1 - \alpha$, the directional error rate is controlled simultaneously with the FWE. The above representation extends readily to any set of hypotheses satisfying the free combinations condition or involving fixed sequence testing schemes. In other situations the integration regions might be more laborious to determine.

As an application we have evaluated the CERs for the comparison of successive means for all combinations of $\nu \in \{1, 2, 4, 8\}$, $\delta_j = 0(0.5)8$ for some $j \leq g$, $g = 3(1)8$ and 5 different correlation patterns. The maximum CER out of the 2040 investigated situations was 0.050127, which lied within the integration tolerance error. Hence, no case of CER was found, which is in accordance with the results of the broad study from Westfall et al. (2000).

Figure 3 depicts the typical behaviour of the CER. Starting from a fixed value $\alpha_0 < \alpha$, the CER first decreases with increasing noncentrality parameter. After achieving its minimum
value, the curve asymptotically converges to the $\alpha$ bound from below, without intersecting it, however. A proof of the limiting behavior is given by Westfall et al. (2000). We observed this behaviour in all parameter configurations under investigation.

Figure 3: Analytical computation of CERs for the comparison of successive means at different values of the noncentrality parameter; $g = 5$, balanced case for different degrees of freedom: $\nu = 1$ (solid line), $\nu = 2$ (dotted line), $\nu = 4$ (dense dashes), $\nu = 8$ (sparse dashes).

4.3 Predictive power

In a usual power analysis, the probability of correctly rejecting the null hypothesis is calculated for a specified deviation from $H_0$, conditional on a certain set of noncentrality parameters. However, it may be more natural to take into account the a priori uncertainty of the noncentrality parameters when calculating power, and compute “predictive power” rather than conditional power (see e.g., Spiegelhalter and Freedman, 1986; Westfall, Krishen and Young, 1997).

Suppose preliminary data $y$ yield a posterior distribution $\theta|y \sim N_k(\theta_0, \sigma^2 A)$, as shown in Carlin and Louis (2000, p. 35). Assuming the (unknown) $\sigma^2$ to be the same in both studies, the current data satisfy $\hat{\theta}|\theta \sim N_k(\theta, \sigma^2 V)$. Predictive power may be calculated using the unconditional (on $\theta$) distribution of $\hat{\theta}$, which is $\hat{\theta} \sim N_k(\theta_0, \sigma^2 W)$, where $W = A + V$. 

19
The integrations are only slightly more complicated in this setup. For example, the predictive probability \( P(\hat{\theta}_i / \{s^2 v_{ii}\}^{1/2} > c) \) is given by \( P(T_i > c(v_{ii}/w_{ii})^{1/2}) \), where \( T_1, \ldots, T_k \) is multivariate t with dispersion matrix \( R_i = (\text{diag}\{W\})^{-1/2} W (\text{diag}\{W\})^{-1/2} \), with degrees of freedom from the current study, and with noncentrality vector \( \delta \) having elements \( \delta_i = \theta_{0i}/\{\sigma^2 w_{ii}\}^{1/2} \).

5 Numerical Example

Probably the most common clinical trials involving more than two treatment groups are dose-response studies. For the estimation of a minimum effective dose (MED) the preceding results are applied to a fixed sequence testing scheme involving one-sided non-pairwise maxT based statistics. Consider the hypotheses \( H_j : \mu_1 = \ldots = \mu_j \) against \( K_j : \mu_1 \leq \ldots \leq \mu_j, \mu_1 < \mu_j \) for \( j = 2, \ldots, g \). Starting with \( j = g \), the hypotheses are tested until a non-significant result is obtained or \( j = 2 \). In either case the procedure stops and the MED is estimated to be the lowest \( j \), such that \( H_j, \ldots, H_g \) are all rejected.

We re-analyze the example given in Ruberg (1989) in terms of power calculation. The effect of a new drug was measured by an increase in the weight of a particular organ in mice. The study consisted of a control group and four dose groups with \( n = 12 \) replications per group. The mean responses were 6.20, 6.14, 6.54, 7.67 and 9.37 for the \( g = 5 \) treatment groups. For these data, the pooled variance estimate \( \hat{\sigma}^2 = 6.33 \) with \( \nu = 55 \). For the test statistics we use the step contrasts \( T_j = \max \left\{ C_j' \bar{X} / \sqrt{\hat{\sigma}^2 C_j' DC_j} \right\}, j = 2, \ldots, g \), where
\( \bar{X} = (\bar{X}_1, \ldots, \bar{X}_g) \) is the vector of the usual arithmetic means and \( D = \text{diag}(n^{-1}, \ldots, n^{-1}) \).

In the present balanced case the contrast matrix \( C'_j \) has the elements \( c_{li} = i - j \) if \( l \leq i \), \( c_{li} = i \) if \( i < l \leq j \) and \( c_{li} = 0 \) otherwise, \( i = 1, \ldots, j - 1, l = 1, \ldots, g \). These contrasts have been introduced independently by Hirotsu (1979) and Bauer and Hackl (1985) and they are particularly powerful for single-inequality alternatives.

Assume that the true MED among the investigated doses is expected to be the second highest dose group. For a valid comparison to predictive power values, we compute the noncentrality parameters based on the mean responses of the first study and assume the same sample sizes for the second study. Thus, \( m' = 2 \), and \( \delta_5 = (1.51, 2.55, 3.36, 3.37) \) and \( \delta_4 = (0.70, 1.29, 1.64) \), where \( \delta_j \) denotes the noncentrality vector at stage \( j, j = 4, 5 \). Moreover, we set \( \alpha = 0.05 \) and the integration error \( \varepsilon = 0.0001 \). The complete power allows the assessment of the sensitivity to estimate correctly the true MED after conducting the study. Hence,

\[
\Pi_{4,5}^{\text{compl}} = P(T_5 > c_5 \land T_4 > c_4) \\
= 1 - P(T_5 \leq c_5) - P(T_4 \leq c_4) + P(T_5 \leq c_5 \land T_4 \leq c_4) \\
= 1 - \int_{(-\infty,c_5]^3} \cdots \int_{(-\infty,c_4]^3} f(x)dx \\
- \int_{(-\infty,c_5]^3} \cdots \int_{(-\infty,c_4]^3} f(x)dx + \int_{(-\infty,c_5]^3 \times (-\infty,c_4]^3} f(x)dx.
\]

For these data, \( c_5 = 2.2012, c_4 = 2.1037 \) and \( \Pi_{4,5}^{\text{compl}} = 0.4200 \). In contrast, the individual power for \( H_5 \), for example, is the probability of finding a (global) trend among all groups. Thus, \( \Pi_5^{\text{ind}} = P(T_5 > c_5) = 0.9539 \). Averaging over the individual power values for \( j = 4 \) and \( j = 5 \) yields a value for the expected average probability of correctly estimating the true
MED. Thus, $\Pi_{4,5}^{\text{ave}} = (\Pi_{5}^{\text{ind}} + \Pi_{4}^{\text{ind}})/2 = (\Pi_{5}^{\text{ind}} + \Pi_{4,5}^{\text{compl}})/2 = 0.6869$. The application of suitable iterative procedures would lead to the minimum sample size required to achieve a pre-specified power. Note that if two-sided trend tests were used for the present study, the CER could also be evaluated without much additional effort.

It may be more natural to include the stochastic uncertainty of the prior results via predictive power. Since the same set of inferences is used, it follows that $A = V$ and thus $R_1 = R$. Given that, the only parameters that change are the integration limits and $\delta$. Both parameters change by the same amount, so that the integral values should not differ too much. Indeed, the values for the predictive power are 0.9187 for finding a global trend and 0.5049 for estimating the true MED. While $\delta$ is small, the prior distribution allows that sometimes it can be larger, and this makes the predictive power a little larger. But when $\delta$ is such that power is around 0.8 to 0.9, the predictive power is smaller, because smaller values of the conditional power are possible, making the distribution of possible values of conditional power is skewed to the left).

6 Conclusions

This paper introduces general concepts for the analytical evaluation of operating characteristics (such as power values and CERs) of several classes of multiple testing procedures. Various applications and examples demonstrate the wide range of application of these results. In fact, all maxT based statistics within CTPs, fixed sequence tests and intersection-union
tests fall within the scope of this article. The differences between individual, average, minimal and complete power are discussed in detail with emphasis on the derivation of the particular integration regions. Correct side decisions at the different steps may be taken into account.

Two different ways of determining the appropriate integration regions are introduced. The first method of decomposing the rejection region into finite unions of disjoint hyper-rectangles can also be applied to other set-ups than discussed here. The second way of applying union-intersection representation leads to a closed form representation of the regions to be integrated. But its application depends more on the structure of the present problem of interest. In any case, the subsequent analysis using adequate integration routines lead to robust and efficient computations of these integrals. Numerical software for the computation of the multivariate $t$ integrals as well as dynamic algorithms involving the inclusion-exclusion formulas are available from the authors.

The application of all-pairwise comparisons provides potentially the most complicated integration regions. Other practically relevant cases as many-to-comparisons, comparisons of successive means or fixed sequence based approaches require usually much less computational effort. Due to the lower number of intersection hypotheses and possible numerical integration short cuts, the power evaluation is also feasible for higher number of treatments, say $g < 6$. Straightforward modifications of the algorithms also allow for a simultaneous computation of several power types for a given expectation vector. In cases where such expectations are difficult to determine in advance, the predictive power approach is a possible alternative.
The results of this paper facilitate a comparison between CTPs with other stepwise methods. In particular, the results help to quantify the power increase of stepwise methods over single-step procedures. Since the latter procedures can be regarded as a global test within a CTP, the integral representations give a complete description about the additional size of the critical region when using stepwise methods. Moreover, sample size determination in designed experiments (a crucial task for any clinical trial) might find the results useful. Such sample size determinations do not only apply to the multiple comparison of several treatment groups, but they also apply to the estimation of certain parameters of interest, such as the MED in dose-response studies. Finally, precise computations of CERs alleviates the ongoing concern about making one-sided decisions after two-sided tests.

REFERENCES


*of Statistical Application Research* 26 12–21.


