

Estimating Directional Error Rates of Step-wise Multiple Comparisons Methods Using Distributed SAS Computing and Variance Reduction

Peter H. Westfall[†], Randall D. Tobias[‡], and Frank Bretz[§]

[†]Texas Tech University

[‡] SAS Institute Inc.

[§] University of Hannover

December 7, 2000

Running head: Directional Errors

Keywords: closure principle, multiple testing, FWE, logical constraints, Type III errors

Address for correspondence:

Peter Westfall
Department of ISQS
Texas Tech University
Lubbock, TX 79423.
phone: (806) 742 2174
FAX: (806) 742 2099
e-mail: westfall@ba.ttu.edu

Abstract

Directional errors are investigated for stepwise multiple testing procedures using numerical and Monte Carlo methods. We develop a Monte Carlo variance reduction method amenable to this purpose. A factorial design is developed to identify possible problem areas, and the asynchronous processing facilities of SAS/Connect[®] software are employed to simulate directional error rates using distributed computing. No cases of excess directional error are found for typical applications involving noncentral multivariate t distributions.

Introduction

We consider testing multiple null hypotheses $H_i : \theta_i = 0$ versus their two-sided alternatives $G_i : \theta_i \neq 0$, $i = 1, \dots, k$ using methods that control the familywise error rate (FWE). To define FWE, 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 $\text{FWE} = P(A_1)$. Multiple testing methods commonly aim to control FWE at or below pre-specified level α (Hochberg and Tamhane, 1987).

Having rejected H_i , one naturally wishes to claim that the sign of θ_i is the same as that of its estimate $\hat{\theta}_i$. 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 in the literature) is defined as a rejection of a false null, but where the sign of θ_i is opposite that of $\hat{\theta}_i$. Let A_2 be the event that there is at least one sign error among the true non-null effects,

and define the combined error rate (CER) as $\text{CER} = P(A_1 \cup A_2)$.

Stepwise testing methods are powerful methods for controlling FWE (Marcus, Peritz and Gabriel, 1976; Grechanovsky and Hochberg, 1999), but do not necessarily control CER. Shaffer (1980) gave a counterexample involving shifted Cauchy distributions; however, she also noted that for independent tests statistics satisfying certain distributional conditions (which include the normal but rule out the Cauchy case), CER is controlled by at least one stepwise method, Holm’s (1979a) step-down method. Holm (1979b) also noted CER control for conditionally independent tests including noncentral multivariate t with identity dispersion matrix. Finner (1999) further extended the class of stepwise tests that control CER to include some step-up tests, closed F tests, and modified S -method tests. He also noted that, while specialized procedures guaranteeing CER control have been developed (e.g., Bauer et al., 1986), they are often less powerful than standard closed and stepwise tests. CER control for procedures based on noncentral multivariate t with correlations, with particular applications being many-to-one and all pairwise comparisons, remains an open question.

The purpose of this article is to systematically explore CER rates of stepwise testing methods relevant to ANOVA studies involving correlated comparisons, using both analytic and simulation-based methods. We develop a Monte Carlo variance reduction method for CER estimation, and apply it using a large distributed computing system.

The Stepwise Tests

The simplest of stepwise tests is Holm's (1979a) method. Let the ordered p -values $p_{(1)} \leq \dots \leq p_{(k)}$ correspond to hypotheses $H_{(1)}, \dots, H_{(k)}$. Define

$$\tilde{p}_{(i)}^{(B)} = (k - i + 1)p_{(i)},$$

which is an unadjusted p -value for the joint hypothesis $\cap_{j=i}^k H_{(j)}$ based on the minP test statistic and the Bonferroni inequality. If $h \in \{1, \dots, k\}$ is the least index such that $p_{(h)}^{(B)} > \alpha$, then Holm's method consists of rejecting $H_{(1)}, \dots, H_{(h-1)}$ and not rejecting $H_{(h)}, \dots, H_{(k)}$.

Formally, Holm's step-down algorithm is employed as follows:

Set $h \leftarrow 1$.

While $(p_{(h)}^{(B)} \leq \alpha)$, $\{h \leftarrow h + 1\}$.

Reject $H_{(1)}, \dots, H_{(h-1)}$; do not reject $H_{(h)}, \dots, H_{(k)}$.

The Holm procedure is a closed testing procedure (Marcus et al., 1976; Grechanovsky and Hochberg, 1999) that uses the Bonferroni minP test statistic (or equivalently, the maxT statistic commonly used in the Tukey and Dunnett range tests) for each intersection hypothesis. There are at least two ways to improve this method: (i) use the exact distribution of minP (or maxT) for each intersection, rather than the conservative Bonferroni test, and (ii) in cases where the hypotheses are logically constrained, one can reduce the Bonferroni multiplier $(k - i + 1)$, further improving power.

To improve upon Holm's method using exact distributions, one may calculate the p -values for the intersection hypotheses using the actual distribution of minP, instead of using the

Bonferroni bound $\tilde{p}_{(i)}^{(B)} = (k - i + 1)p_{(i)}$. Define

$$\tilde{p}_{(i)}^{(E)} = P(\min_{l \in \{r_i, \dots, r_k\}} P_l \leq p_{(i)} \mid \cap_{j=i}^k H_{(j)}),$$

where r_i is the index of the i th ordered p -value (Westfall and Young, 1993, p. 66). The stepwise testing algorithm then proceeds exactly as with the Holm algorithm, but using the joint p -values $\tilde{p}_{(i)}^{(E)}$ instead of $\tilde{p}_{(i)}^{(B)}$. When the hypotheses are comparisons with a control, this method is equivalent to a Dunnett's stepdown method, see Naik (1975), Marcus et al. (1976), and Dunnett and Tamhane (1991).

Various authors have described improvements for the case where the hypotheses are logically constrained. Such constraints occur when the truth of intersections of particular nulls force the truth of other nulls, as happens in the case of all pairwise comparisons of group means. The so-called REGWQ method of PROC GLM in SAS/STAT[®] (see Tamhane, 1996, and references contained therein) is one such method. Shaffer (1986) shows how to incorporate restrictions using the Bonferroni minP test for more general comparisons, and Westfall (1997) improved upon Shaffer's method by using the actual distribution of minP rather than the Bonferroni minP test. We use the Westfall (1997) (hereafter W97) method because it always finds at least as many rejections as the Holm, REGWQ, and Shaffer methods, and is essentially equivalent to Dunnett and Tamhane's (1991) step-down method. If no excess CER is found using W97, then none can be found using Holm, Shaffer, REGWQ, or step-down Dunnett either.

Specifically, the W97 method is identical to the Holm step-down method with exact minP tests using $\tilde{p}_{(i)}^{(E)}$, but uses instead the logically constrained joint p -values

$$\tilde{p}_{(i)}^W = \max_{K \in S_i} \Pr(\min_{l \in K} P_l \leq p_{(i)} \mid \cap_{j \in K} H_j),$$

which is the maximum of probabilities that the minP statistic is less than the observed p at stage i , with maximum being taken over all sets of hypotheses that possibly can be true at the given stage, given that all previously rejected hypotheses are false.

The joint p -values $\tilde{p}_{(i)}^W$ and $\tilde{p}_{(i)}^E$ can in some cases be computed exactly (e.g., when they are related to the Dunnett or studentized range distributions), but otherwise may require Monte Carlo methods. Westfall (1997) describes a simple and efficient method involving Monte Carlo variance reduction method for doing so. Software for the implementing W97 method using variance reduction is readily available, see Westfall et al. (1999).

Numerical Evaluation of Directional Errors; Shaffer's Counterexample

Shaffer (1980) found a counterexample showing where the Holm step-down method using exact p -values $\tilde{p}_{(i)}^{(E)}$ does not control the CER: let V_1 and V_2 be independent t -distributed random variables with ν degrees of freedom, and define test statistics $T_i = V_i + \theta_i$, $i = 1, 2$. Suppose that $\theta_1 > 0$ and $\theta_2 = 0$. The Holm exact step-down testing method requires the distribution of minP, which may be equivalently obtained using the null distribution

of $\max(|T_1|, |T_2|)$. With $\nu = 1$, the 95th percentile of this distribution is 25.13 and the 95th percentile of the null distribution of $|T_i|$ is 12.71, but more generally, these quantities are denoted $t_{\nu,05}^{(2)}$ and $t_{\nu,05}^{(1)}$ respectively. Thus, the hypotheses are rejected according the following algorithm, which also includes directional determinations.

1. Let $T_{(1)} = \max(|T_1|, |T_2|)$. If $T_{(1)} > t_{\nu,05}^{(2)}$, then reject $H_{(1)}$ and classify the sign of $\theta_{(1)}$ according to the sign of the corresponding T_i , and continue. Otherwise do not reject $H_{(1)}$ or $H_{(2)}$ and stop.
2. Let $T_{(2)} = \min(|T_1|, |T_2|)$. If $T_{(2)} > t_{\nu,05}^{(1)}$, then reject $H_{(2)}$ and classify the sign of $\theta_{(2)}$ according to the sign of the corresponding T_i . Otherwise do not reject $H_{(2)}$.

Figure 1 shows the region for which no type I error and no sign error occurs when $\nu = 1$. In this example, CER is readily computed analytically as the probability of the complement of the shaded region, shown in Figure 2 for varying θ_1 and ν (for $\nu > 1$, the region is smaller than shown in Figure 1). Note that CER is not controlled at the 5% level, although the excess is only slight. Shaffer's counterexample specifically used the Cauchy distribution ($\nu = 1$); CERs slightly in excess of .05 are seen for other shifted t -distributions as well, but the excess rapidly diminishes as ν increases.

The shifted t distributions used in Figure 2 are not typically seen in applications. Non-central t distributions are more common. There is a subtle but important distinction between shifted and noncentral t 's that allows CER control for the latter. Finner (1999) gives sufficient conditions on distributions so that CER is controlled; such distributions include

noncentral t with identity dispersion matrix.

The bivariate noncentral t distribution is defined as follows: let (V_1, V_2) be standard normal variates with correlation ρ , and let U be independently distributed as chi-square with ν degrees of freedom. Letting $T_i = (V_i + \theta_i)/\sqrt{U/\nu}$, the vector (T_1, T_2) is distributed as bivariate t with correlation ρ , degrees of freedom ν , and noncentrality vector (θ_1, θ_2) . This distribution often arises naturally, with known value of ρ , e.g. when comparing the means of two treatment groups against a common control. Critical values for multiple comparisons using the null distribution of (T_1, T_2) are readily obtained from available statistical software (e.g., PROBMC of SAS[®]). Defining $t_{\nu, \rho, .05}^{(2)}$ to be the 95th percentile of the distribution of $\max(|T_1|, |T_2|)$, Holm's exact step-down testing method with directional determinations is given exactly as above, but with $t_{\nu, \rho, .05}^{(2)}$ replacing $t_{\nu, .05}^{(2)}$. CER can then be calculated by integrating the noncentral bivariate t distribution analytically over a region similar to that shown in Figure 1. To integrate this distribution, we used the intrinsic bivariate normal probability function PROBBNRML of SAS[®], and numerically integrated over the distribution of U so that the integral was one-dimensional. Figures 3a-3d display CER estimates as a function of θ_1 and ρ for $\nu = 1, 2, 4$ and 8. In no case is CER in excess of .05 found. That CER is controlled for $\rho = 0$ has already been proven; but there does not yet exist a general proof of CER control for $\rho \neq 0$.

The Variance Reduction Method

The analytic results of the previous section show that there are no excess CERs in simple cases involving noncentral multivariate t tests and $k = 2$. To evaluate CER in more complex studies requires Monte Carlo methods, first to incorporate the correlations correctly for the $\tilde{p}_{(i)}^W$ p -values, and second to estimate the proportion of samples that yield one or more (type I or sign) errors. As suggested by Figure 2, any excess CER that may be found is likely to be small, and therefore a large and efficient simulation study will be needed to detect a problem. Since we must simulate a simulation-based method and obtain accurate results, we require efficient estimation procedures and extensive computing capabilities. In this section we describe an efficient Monte Carlo procedure to estimate CER.

With simple Monte Carlo (SMC) simulation, one specifies the model, simulates the tests, and notes whether a type I or sign error occurred in the sample. This process is repeated NSIM times, the estimated CER is \hat{p}_{SMC} , the percentage of the NSIM samples showing a type I or sign error, with Monte Carlo standard error $\sqrt{\hat{p}_{SMC}(1 - \hat{p}_{SMC})/NSIM}$.

We propose a combination of Simple Monte Carlo (SMC) and Control Variate Monte Carlo (CV) to improve the accuracy of the estimated CER.

The condition we are counting as an error is identified by an indicator variable V , with $V = 0$ when there is no type I and no sign error, and with $V = 1$ otherwise. Thus, $CER = E(V)$. The CV method expresses $E(V)$ as $E\{W + (V - W)\} = E(W) + E(V - W)$, where W is a control variate for which $E(W)$ is known, and where $V - W$ ideally has small

variance. Thus, the CV estimate of CER is $E(W) +$ (the average of the NSIM values $V - W$).

To define the control variate W , let T_{0i} be the t -statistic for the i th test that is truly null, and T_{1i} be the i th t -statistic for a non-null test. Assume, without loss of generality, that directional errors occur when T_{1i} is positive and significant (otherwise, just use $-T_{1i}$).

Define $W = \sum I(|T_{0i}| > t_{\alpha/(2k)}) + \sum I(T_{1i} > t_{\alpha/(2k)})$, where “ I ” is the indicator function. (Note that W is the total number of type I and sign errors that occur when the simple Bonferroni method is used.) Then $E(W) = \alpha m/k + \sum P(T_{1i} > t_{\alpha/(2k)})$, where $m =$ number of true nulls, and where the latter summand can be calculated exactly using the noncentral t -distribution. The control variate estimate, \hat{p}_{CV} , is then the analytic expression for $E(W)$ plus the simulated average of the $V - W$'s.

The estimated CER is improved by combining \hat{p}_{SMC} and \hat{p}_{CV} using generalized least squares (GLS). Letting \hat{V} be the sample covariance matrix of the NSIM pairs $(V, V - W)$, the GLS estimate is

$$\hat{p}_{GLS} = (\hat{p}_{SMC} \quad \hat{p}_{CV}) \hat{V}^{-1} \mathbf{1} / \mathbf{1}' \hat{V}^{-1} \mathbf{1},$$

where “ $\mathbf{1}$ ” is the (2×1) vector of 1's. The GLS estimate has uniformly smaller variance than either the SMC or CV estimates for large NSIM.

The Distributed Computing Environment

The simulations were run on a distributed system of 68 different HP 9000 Unix workstations, all running the SAS System, Version 8.2 for the UNIX operating system, with clock speeds

of either 236 MHz or 400 MHz. These servers were shared with other users, so the effective clock speed was somewhat less than this. The simulations were coordinated by a SAS macro program running on a single client machine which used the multiprocessing capabilities of SAS/CONNECT[®] to run the simulations in parallel as follows:

1. The client launched SAS sessions on each of the 68 servers, and looped through each scenario to be simulated.
2. For each scenario, the client asynchronously executed the simulation in groups of 250 replicates on the servers. SAS/CONNECT makes it possible for the client to monitor completion of jobs on the servers. Thus, as soon as a server completed one 250-replicate simulation, another group was launched on it. This (crude) form of load-balancing greatly increased the scaling efficiency of the distributed computation.
3. Finally, when a total of 100,000 replicates were completed, the client collated the results and computed the variance reduced estimate of CER.

Using this distributed processing scheme, simulations that would have taken 3 years' worth of computing were accomplished in about a month's worth of desktop time.

Design of the Simulation Study

In all cases we assume a one-way ANOVA with possibly unbalanced sample sizes: $y_{ij} = \mu_i + \epsilon_{ij}$; $i = 1, \dots, g$, $j = 1, \dots, n_i$, where the ϵ_{ij} are an i.i.d. sample from $N(0, \sigma^2)$,

and where the μ_i and σ^2 are fixed, unknown constants. Since it is known that the W97 method controls FWE, simulations must use alternative configurations to detect possible excess CER. Directional errors can occur only in one tail, thus the contribution to CER tends to be greatest from the true type I errors. If no nulls are true, then the CER is substantially smaller (Bofinger, 1985), thus we chose to put the nonzero effect in only one group, the last one.

We consider four simulation design factors, “COMPARE,” “NGROUPS,” “EFFSIZE,” and “NCONFIG,” defined as follows:

- COMPARE refers to set of contrasts considered, with seven levels: AllPairs, Controlg, Control1, Seq, Ave, AltAve, and AltWgtAve.
 - “AllPairs” refers to all pairwise comparisons.
 - “Controlg” refers to pairwise comparisons against the last group.
 - “Control1” refers to pairwise comparisons against the first group.
 - “Seq” refers to sequential comparisons $\mu_i - \mu_{i+1}$, $i = 1, \dots, g - 1$.
 - “Ave” refers to comparisons of an individual mean with the average of all other means, having the form $\mu_i - \sum_{j \neq i} \mu_j / (g - 1)$.
 - “AltAve” is like “Ave,” but where the average uses alternating signs.
 - “AltWgtAve” is like “AltAve,” but with varying weights.
- NGROUPS is g , at 10 levels: $g = 4, \dots, 13$. For COMPARE=AllPairs, only $g \leq 6$ was

considered to save time.

- EFFSIZE refers to the value of the nonzero group mean μ_g relative to the square-root of its sample size, at five levels: $0.025/\sqrt{n_g}$, $0.05/\sqrt{n_g}$, $0.1/\sqrt{n_g}$, $0.3/\sqrt{n_g}$, and $1/\sqrt{n_g}$. Most of the runs used levels $0.1/\sqrt{n_g}$, $0.3/\sqrt{n_g}$, and $1/\sqrt{n_g}$; the additional levels $0.025/\sqrt{n_g}$, $0.05/\sqrt{n_g}$ were used with COMPARE=Seq and $g > 7$ as these appeared to be the most likely places to discover excess CERs.
- NCONFIG refers to sample size configuration, at five levels, $(1, 1, 1, \dots, 1, 2)$, $(1, 2, 1, \dots, 1, 1)$, $(2, 1, 1, \dots, 1, 1)$, $(2, 2, 2, \dots, 2, 2)$, and $(4, 4, 4, \dots, 4, 4)$

In all, 432 combinations of these design parameters were considered, although not in a complete factorial arrangement. The “AllPairs” simulations are quite time-consuming for large g , so we performed only a few “AllPairs” simulations at $g = 6$ that seemed most likely to exhibit liberal CERs. Additional runs were also made for the “Control1” cases since it seemed to be the case with largest CER. The “Seq” comparisons also exhibited relatively larger CERs. These comparisons are less time-consuming than “AllPairs,” so we considered “Seq” comparisons all the way to $g = 13$, with the additional smaller effect sizes as noted above.

The main simulation used NSIM=100,000 data sets for each configuration. In all cases we took the simulation sample size for the W97 method to be NSAMP= 499 (this is the inner loop of the simulation), but made the outer loop simulation size much larger to reduce the major source of Monte Carlo error (see e.g. Hope, 1968).

Results of the Simulation Study

Table 1 displays estimated CERs with 95% (unadjusted) confidence intervals and design configurations for the cases with largest CERs. In no cases were the estimates larger than .05, although in the cases shown, the confidence interval extends beyond .05. Thus, in general our study uncovered no convincing evidence of CER's in excess of 0.05 for any of the cases we simulated.

We also evaluate LS-Means for the various design parameters, shown in Tables 2a.-2c. For the "COMPARE" factor, CER is largest for Control1 and Seq, which is not surprising since these configurations have the greatest percentage of tests that are true nulls. CER is relatively less affected by NGROUPS. The largest CER are seen for small effect sizes, which is where directional errors are most likely. Finally, the CER appears largest for the $(1, 1, 1, \dots, 1, 2)$, $(1, 2, 1, \dots, 1, 1)$, and $(2, 1, 1, \dots, 1, 1)$, where the marginal t -statistics are distributed as Cauchy, corroborating Figure 3(a).

LS-Means of efficiencies for the proposed GLS estimation method relative to simple Monte Carlo are shown in Table 3. The COMPARE and NCONFIG effects exhibited significant variation in efficiencies. Large gains are possible for larger degrees of freedom, evidenced by the larger efficiencies for the $(2, 2, \dots, 2)$ and $(4, 4, \dots, 4)$ levels of the NCONFIG factor. Note that an efficiency of 3.0 means that the simulation sample size (NSIM) for simple Monte Carlo would have to be 300,000 to achieve identical precision as that of the GLS method. Note also that the efficiency is always greater than 1.0, and was larger than 7.0 in several

cases.

Conclusion

In this article we have investigated CER for common statistical models using for a broad class of stepwise testing methods, and have found no evidence of CER in excess of .05. A large simulation study was performed, one that would have taken three years of desktop computing, but which took only a month with distributed computing. A Monte Carlo variance reduction technique was developed specifically for CER estimation to improve the efficiency.

While it remains to be proven that CER is controlled mathematically for the entire class of stepwise procedures and models considered in this article (particularly, the case of correlated estimates), our results suggest that practicing statisticians may feel confident in making directional determinations when using such methods.

ACKNOWLEDGEMENT

The authors wish to thank Juliet Shaffer for helpful comments.

REFERENCES

Bauer, P., Hackl, P., Hommel, G., and Sonnemann, E. (1986). Multiple testing of pairs of one-sided hypotheses. *Metrika* **33** 121–127.

Bofinger, E. (1985). Multiple comparisons and type III errors. *J. Amer. Statist. Assoc.* **80** 433–437.

Dunnett, C.W. and Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statist. Medicine* **11**, 1057–1063.

Finner, H. (1999) Stepwise multiple test procedures and control of directional error rates. *Annals of Statistics* **27**: 274–289.

Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* **76** 79–91.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley, New York.

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

Holm, S. (1979b) A Stagewise directional test based on t Statistics. Statistical Research Report 1979-3, Inst. Mathematics, Chalmers Univ. Technology, Gothenberg.

Hope, A.C.A. (1968) A simplified Monte Carlo significance test procedure. *J. Royal. Statist. Soc.* **30** 582–598.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660.

Naik, U.D. (1975). Some selection rules for comparing P processes with a standard. *Commun. Statist. Ser. A* **4**, 519–535.

Shaffer, J.P. (1980) Control of directional errors with stagewise multiple test procedures. *Annals of Statistics* **8**: 1342–1347.

Shaffer, J.P. (1986) Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

Tamhane, A.C. (1996). Multiple Comparisons. S. Ghosh and C.R. Rao, eds., *Handbook of Statistics*, Vol. 13, 587–630.

Westfall, P.H., and Young, S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* Wiley, New York.

Westfall, P.H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Statist. Assoc.* **92** 299–306.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS[®] System*, Cary, NC: SAS Institute Inc.

Table 1: Design points with largest CERs.

CER	95% CI for CER	COMPARE	NGROUPS	EFFSIZE	CONFIG
0.04979	(0.04856, 0.05103)	Seq	10	0.025	(1,1,1,....,1,2)
0.04974	(0.04848, 0.05100)	Seq	13	0.05	(1,1,1,....,1,2)
0.04971	(0.04851, 0.05091)	Seq	8	0.05	(1,1,1,....,1,2)
0.04969	(0.04846, 0.05092)	Seq	10	0.1	(1,1,1,....,1,2)
0.04959	(0.04835, 0.05083)	Seq	11	0.1	(1,1,1,....,1,2)
0.04957	(0.04831, 0.05082)	Seq	13	0.1	(1,1,1,....,1,2)
0.04956	(0.04832, 0.05081)	Seq	11	0.05	(1,1,1,....,1,2)
0.04954	(0.04831, 0.05077)	Seq	10	0.05	(1,1,1,....,1,2)
0.04948	(0.04823, 0.05074)	Seq	13	0.025	(1,1,1,....,1,2)
0.04942	(0.04822, 0.05061)	Seq	8	0.025	(1,1,1,....,1,2)
0.04938	(0.04814, 0.05062)	Seq	11	0.025	(1,1,1,....,1,2)
0.04937	(0.04823, 0.05051)	Control1	6	0.1	(1,1,1,....,1,2)
0.04935	(0.04815, 0.05055)	Seq	8	0.1	(1,1,1,....,1,2)
0.04933	(0.04815, 0.05050)	Seq	7	0.1	(2,1,1,....,1,1)
0.04932	(0.04818, 0.05046)	Control1	6	0.3	(1,1,1,....,1,2)
0.04915	(0.04798, 0.05032)	Seq	7	0.3	(2,1,1,....,1,1)
0.04911	(0.04797, 0.05024)	Seq	6	0.1	(1,1,1,....,1,2)
0.04900	(0.04776, 0.05024)	Seq	12	0.05	(1,1,1,....,1,2)
0.04896	(0.04769, 0.05023)	AllPairs	6	0.1	(1,1,1,....,1,2)
0.04892	(0.04779, 0.05006)	Seq	6	0.1	(1,2,1,....,1,1)

Table 2. LS-Means of CER for the various design parameters.

<u>COMPARE</u>	<u>LS-Mean</u>	<u>NGROUPS</u>	<u>LS-Mean</u>	<u>EFFSIZE</u>	<u>LS-Mean</u>	<u>NCONFIG</u>	<u>LS-Mean</u>
AllPairs	0.043	4	0.035	0.025	0.036	(1,1,1,...,1,2)	0.040
Controlg	0.022	5	0.037	0.05	0.036	(1,2,1,...,1,1)	0.039
Control1	0.045	6	0.038	0.1	0.036	(2,1,1,...,1,1)	0.038
Seq	0.045	7	0.038	0.3	0.035	(2,2,2,...,2,2)	0.031
Ave	0.032	8	0.034	1	0.032	(4,4,4,...,4,4)	0.028
AltAve	0.029	9	0.033				
AltWgtAve	0.030	10	0.034				
		11	0.034				
		12	0.033				
		13	0.034				

Table 3. LS-Means, min and max of efficiency for the design parameters.

COMPARE	LS-Mean	Min	Max	NGROUPS	LS-Mean	Min	Max
AllPairs	1.24	1.11	2.34	4	2.09	1.00	4.44
Controlg	1.72	1.26	3.90	5	1.92	1.00	4.98
Control1	1.79	1.30	3.86	6	1.91	1.00	5.78
Seq	2.04	1.14	4.44	7	1.90	1.00	7.25
Ave	1.95	1.00	7.25	8	1.62	1.25	1.26
AltAve	2.00	1.00	5.91	9	1.59	1.23	1.23
AltWgtAve	1.27	1.00	3.78	10	1.56	1.20	1.20
				11	1.54	1.18	1.18
				12	1.52	1.16	1.16
				13	1.51	1.14	1.15

EFFSIZE	LS-Mean	Min	Max	NCONFIG	LS-Mean	Min	Max
0.025	1.69	1.15	1.26	(1,1,1,...,1,2)	1.02	1.00	1.78
0.05	1.72	1.14	1.26	(1,2,1,...,1,1)	1.02	1.00	1.79
0.1	1.71	1.00	7.25	(2,1,1,...,1,1)	1.01	1.00	1.78
0.3	1.72	1.00	6.92	(2,2,2,...,2,2)	2.08	1.52	3.09
1	1.72	1.00	7.00	(4,4,4,...,4,4)	3.44	1.97	7.25

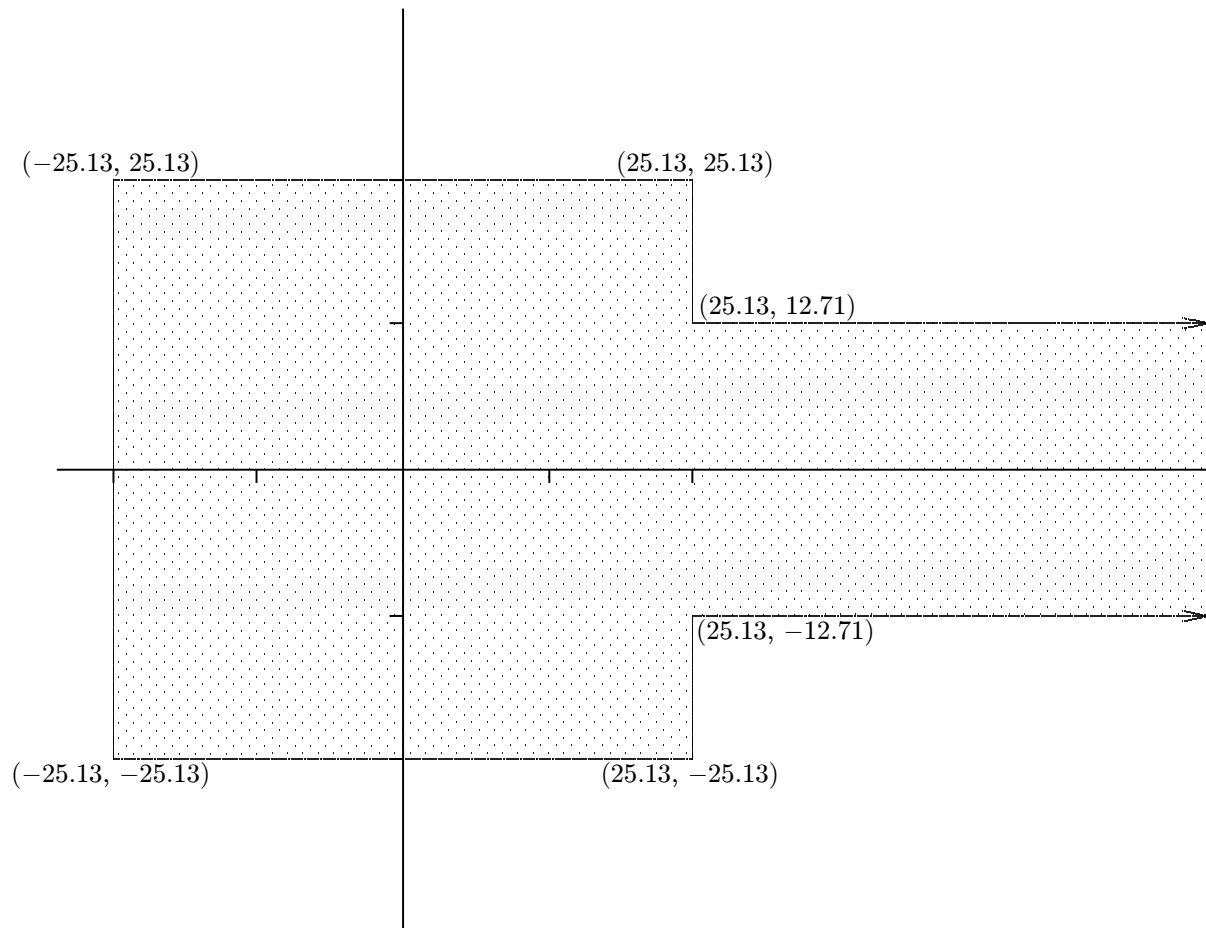


Figure 1. Region of no type I and no sign error with independent Cauchy test statistics T_1, T_2 , when $\theta_1 > 0$, $\theta_2 = 0$.

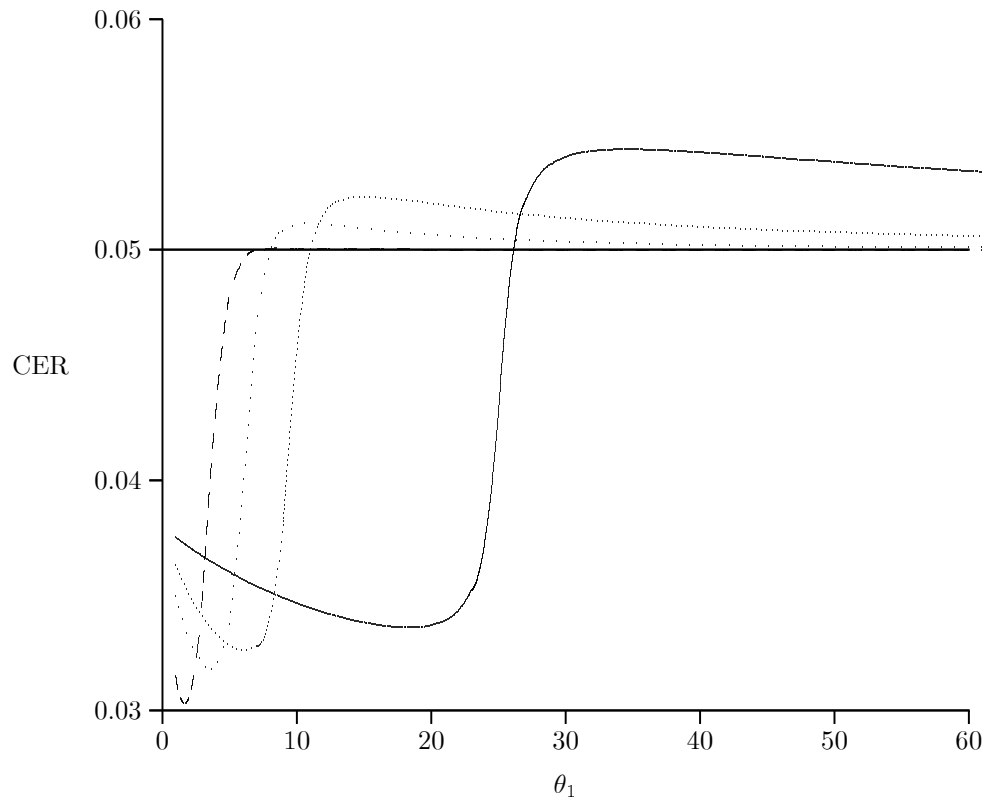


Figure 2. CER as a function of θ_1 with shifted t distributions having $df=1$ (Shaffer's counterexample, solid line), $df=1.5$ (dense dots), $df=2$ (sparse dots), and $df=4$ (dashes).

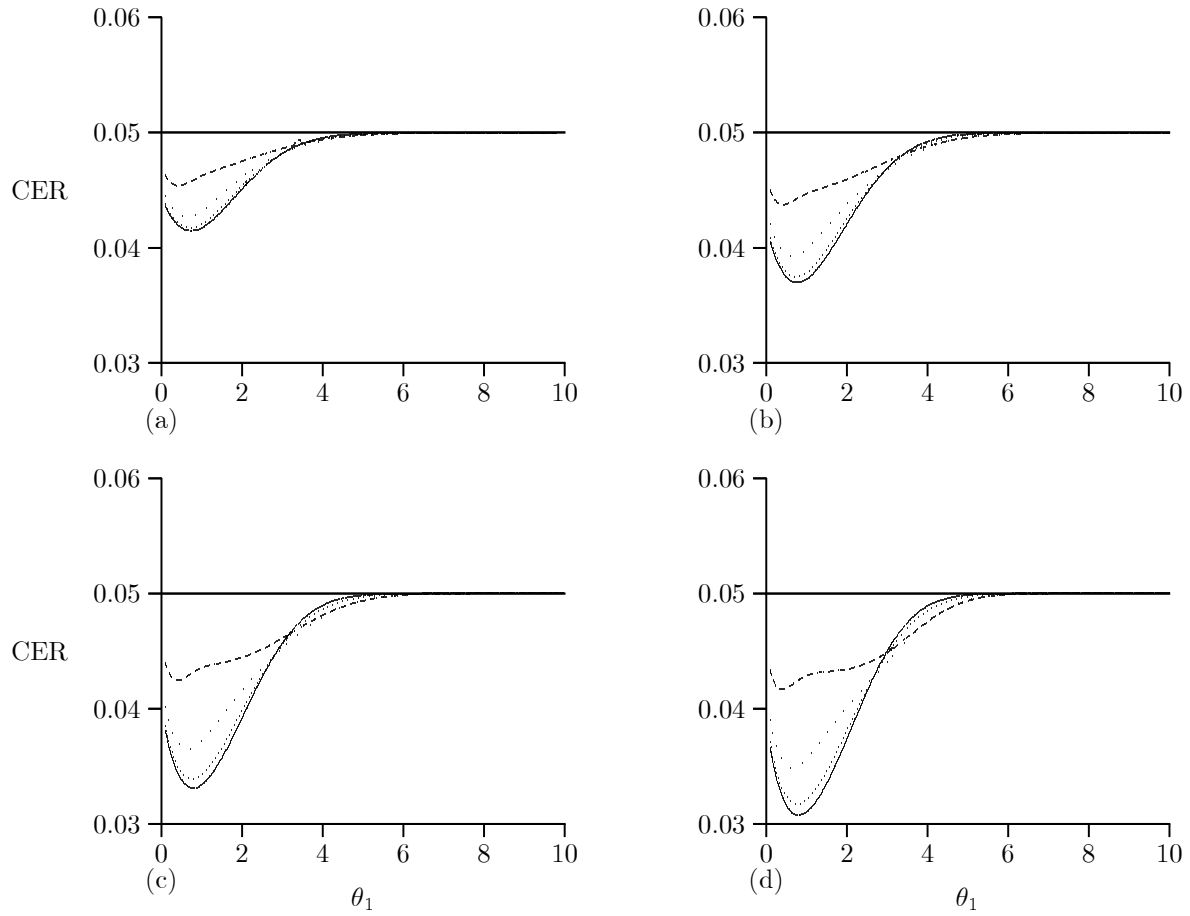


Figure 3. CER as a function of θ_1 with noncentral t distributions with $\rho = 0$ (solid line), $\rho = 0.3$ (dense dots), $\rho = 0.6$ (sparse dots), and $\rho = 0.9$ (dashes). Figure (a) has $\nu = 1$, (b) has $\nu = 2$, (c) has $\nu = 4$, and (d) has $\nu = 8$.