Johns Hopkins University, Dept. of Biostatistics Working Papers

7-23-2008

A NOVEL AND SIMPLE RULE OF THUMB FOR MULTIPLICITY CONTROL IN EQUIVALENCE TESTING USING TWO ONE-SIDED TESTS

Carolyn Lauzon
Department of Biophysics, Johns Hopkins University

Brian S. Caffo
Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health, bcaffo@jhsph.edu

A Novel and Simple Rule of Thumb for Multiplicity Control in Equivalence Testing Using Two One-sided Tests

Carolyn Lauzon and Brian Caffo

Department of Biophysics and Department of Biostatistics

Johns Hopkins University

July 23, 2008

Abstract

Equivalence testing is growing in use in scientific research outside of its traditional role in the drug approval process. Largely due to its ease of use and recommendation from the United States Food and Drug Administration guidance, the most common statistical method for testing (bio)equivalence is the two one-sided tests procedure (TOST). Like classical point-null hypothesis testing, TOST is subject to multiplicity concerns as more comparisons are made. In this manuscript, a condition that bounds the family-wise error rate (FWER) using TOST is given. This condition then leads to a simple solution for controlling the FWER. Specifically, we demonstrate that if all pair-wise comparisons of k independent groups are being evaluated for equivalence, then simply scaling the nominal Type I error rate down by $(k-1)$ is sufficient to maintain the family-wise error rate at the desired value or less. The resulting rule is much less conservative than the equally simple Bonferroni correction. An example of equivalence testing in a non drug-development setting is given.

Keywords: bioequivalence, family-wise error rate, multiple comparisons, t-tests, type I error rate, TOST

1 Introduction

Broadly speaking, scientific research is often thought of as a field that is interested in inductively demonstrating differences between experimental groups while presuming equality under a null (status quo) hypothesis. However, often scientists are not interested in establishing differences, but in proving similarities. It is our experience that questions of similarity or equivalence are as fundamentally important to scientific research as those of differences.

An important example is that of demonstrating the bioequivalence of two drugs, such as an established brand name drug and a new generic equivalent. Bioequivalence refers to establishing a lack of differences in absorption as measured by blood concentration, of two such formulations. Hence, the natural null hypothesis is that the two formulations have different absorption rates on a scale that is biologically relevant. Typically, the metrics being compared are natural logarithms of areas under a plasma/concentration curves obtained by repeated blood samples of subjects having received both drugs in a random order (with a suitable washout period).

We refer to this form of evaluation in the drug approval setting as bioequivalence and reserve the term equivalence for more generic settings. Establishing equivalence generally follows two steps; i) first, a setting-specific meaningful difference in population parameters between two groups is selected and ii) statistical inference is used to establish whether empirical estimates of the parameters are fall within the bounds of the meaningful limits.

Early related work on equivalence testing using symmetric intervals can be found in Westlake (1976). Anderson and Hauck (1983) and Hauck and Anderson (1984) give a more powerful method for a two-way crossover design. Since these early influential

articles, new procedures have been developed based on likelihood intervals (Choi et al., 2007), Bayesian credible intervals (Selwyn and Hall, 1984; Selwyn et al., 1981; Fluehler et al., 1983) and alternative frequentist tests and intervals (Berger and Hsu, 1996; Hsu et al., 1994; Brown et al., 1995), to name a few.

Certainly the most widely used procedure for statistically evaluating equivalence is the two one-sided tests procedure (TOST), which is advocated by the US FDA for establishing bioequivalence. TOST is a form of equivalence testing proposed by Schuirmann (1987). Part of TOST's popularity is that it is theoretically and operationally similar to classical normal-theory hypothesis testing of the equality of population means. Despite their close relationship and the ubiquity of (alternative) research hypotheses of similarity, TOST has been mostly unused in the non-drug development scientific community at large, where classical point-null hypothesis testing of population means is firmly entrenched. A possible reason for the large disparity in usage is not one of utility, but of exposure. Perhaps the greatest evidence supporting this explanation is the frequent misapplication of post-hoc power calculations to data that should be analyzed using equivalence testing (Hoenig and Heisey, 2001; Goodman and Berlin, 1994).

Recently, equivalence testing has made inroads in scientific applications unrelated to drug development (Barnett et al., 2007, 2006). In fact, research papers advocating the use of equivalence testing in a diverse collection of fields have begun to appear (Barker et al., 2002; Tempelman, 2004). We conjecture that as awareness of equivalence testing increases, so will the number of scientists incorporating TOST into their regular statistical toolbox. Hence, it is necessary to develop methods for adapting TOST to the diverse situations scientific data can present.

One example addressed here is that of multiplicity. As in classical hypothesis testing, as more means are compared, the family-wise error rate, the probability of at least one incorrectly rejected null in an family of tests, α_F, rises above that of the set nominal type I error rate, α_N. If enough means are compared, the family-wise error rate becomes un-

acceptably high and must be controlled. Because the foundation for equivalence testing is the same as that of classical hypothesis testing, we look to existing solutions for addressing multiplicity. In order to adapt these solutions a more explicit knowledge of equivalence testing is necessary.

2 Bioequivalence theory

For simplicity, we describe the TOST procedure for comparing two independent group means from normally distributed data, presuming a common variance and equal sample sizes. This setting for equivalence testing has been described in detail elsewhere (Schuirmann, 1987). Briefly, equivalence testing seeks to test if the difference between the two population means, $\Delta\mu$, is within some previously defined tolerance interval $[\theta_l, \theta_u]$. To do this, two sets of disjoint hypotheses are formed. Closely following the description and notation in Schuirmann's original manuscript, we have:

$$\begin{array}{lllll} \text{null hypothesis} & H_{01} : \mu \leq \theta_l & \text{or} & H_{02} : \mu \geq \theta_u \\ \text{alternative hypothesis} & H_{a1} : \mu > \theta_l & \text{and} & H_{a2} : \mu < \theta_u, \end{array} \tag{1}$$

From each pair of hypotheses, test statistics are formed and compared to critical values from Gossett's T distribution. Specifically, H_{01} and H_{02} are rejected if

$$\frac{\Delta\bar{X} - \theta_l}{s\sqrt{2/n}} > t_{df,1-\alpha} \tag{2}$$

and

$$\frac{\theta_u - \Delta\bar{X}}{s\sqrt{2/n}} > t_{df,1-\alpha} \tag{3}$$

respectively; where $\Delta\bar{X}$ is the observed difference in means between the two groups, s is the pooled standard deviation (hence we're assuming a common variance across the two groups), n is the (assumed common) sample size per group, df is the degrees of freedom and $t_{a,b}$ is the b quantile from Gossett's T distribution with a degrees of freedom. The TOST procedure states that if both 2 and 3 are true, then the means are declared equivalent.

Equivalently, test (2) rejects if the lower confidence bound $\Delta\bar{X} - t_{df,1-\alpha_N} s/\sqrt{2/n}$ is above θ_l and test (3) rejects if the upper confidence bound $\Delta\bar{X} + t_{df,1-\alpha_N} s/\sqrt{2/n}$ is below θ_u. Hence, the two one sided test procedure is identical to forming the corresponding $(1-2\alpha_N)$ confidence interval and declaring the two groups equivalent if the interval lies entirely within the tolerance limits. Note that there has been discussion in the literature on whether a $(1-2\alpha_N)$ or $(1-\alpha_N)$ interval should be used. We adopt the former, though emphasize that our conclusions do not depend on this choice.

For simplicity of the discussion, we assume that the toleration limit is symmetrically centered around zero; that is $-\theta_l = \theta_u = \theta$. Then the hypotheses (1) can be restated as

$$H_0 : |\Delta\mu| \geq \theta \quad \text{versus} \quad H_a : |\Delta\mu| < \theta$$

and equations (2) and (3) are restated as

$$\frac{\Delta\bar{X} + \theta}{s\sqrt{2/n}} > t_{df,1-\alpha_N} \text{ and } \frac{\theta - \Delta\bar{X}}{s\sqrt{2/n}} > t_{df,1-\alpha_N}. \tag{4}$$

3 Family-wise error rates for all pair-wise comparisons

A common solution in classical hypothesis testing for handling family-wise error rates is the Bonferroni correction. This correction is widely used because of its simplicity. It is based on the fact that the probability of at least one incorrectly rejected null hypothesis in a collection of tests is bounded by the sum of the probabilities of the individual type I error rates. For example if all pair-wise tests are being performed for k groups, each comparison made with a type I error rate of α_N, then the actual family-wise error rate, α_F follows:

$$\alpha_F \leq \alpha_N \begin{pmatrix} k \\ 2 \end{pmatrix} = \alpha_N k(k-1)/2 \tag{5}$$

Therefore, if α_D is a desired family-wise error rate, then setting $\alpha_N = \frac{2\alpha_D}{k(k-1)}$ bounds the actual family-wise error rate by the desired one. Depending on the setting, this procedure can be very conservative, especially when the outcomes of the tests are correlated. We

further note that the Bonferroni correction applies without modification to equivalence testing; one simply divides the error rate by the number of tests performed.

We seek a less conservative method of multiplicity control. Though relevant research in multiple comparisons (see Giani and Strassburger, 2000; Bofinger, 1985; Giani and Strassburger, 1994; Hsu, 1996) may produce more optimal solutions, our interest lies in simple rules that are easily motivated and implemented.

Under our assumptions, in a standard point null hypothesis, the desired type I error rate is obtained exactly. In equivalence testing using TOST, the null hypothesis includes a range of possible parameter values, and hence the desired type I error rate is obtained only on the boundary of the null parameter space (Schuirmann, 1987). Hence, one attains the type I error rate exactly only when $|\Delta\mu| = \theta$ and even then only as a limit as the effect size tends to infinity. Otherwise, the procedure is conservative.

Consider again the setting where all pair-wise comparisons are being made of k groups, each with a population mean μ_i for $i = 1, \ldots, k$. Without loss of generality, we presume that the means are ordered from least to greatest. If all tests satisfy the null hypothesis, then the means must be at least θ apart. The maximum type I error rate for each comparison is then obtained only when the means are exactly θ apart. That is, $\mu_i - \mu_{i-1} = \theta$ for $i = 2, \ldots, k$. We note that this scenario maximizes the family-wise error rate because: decreasing the length between any two adjacent means renders them equivalent (a violation of the assumption that all null hypotheses are true) while expanding the distances decreases the individual type I error rates (hence decreasing the family-wise error rate).

Note that a Bonferroni correction based on (5) accounts for all possible comparisons, even of the most distal means, which must be at least $k - 1$ times the tolerance limit apart. More specifically, observe that in the most conservative scenario, where the ordered means are exactly θ apart, $(k-1)$ comparisons occur with a true difference $\Delta\mu = 1\theta$, $(k-2)$ comparisons with a true difference $\Delta\mu = 2\theta$ and in general $(k - \ell)$ comparisons are made with a true difference $\Delta\mu = \ell\theta$.

Let α_ℓ be the actual type I error rate for the bioequivalence test performed for a comparison with $\Delta\mu = \ell\theta$. This parameter is derived and discussed in Appendix A. Under this setting, α_1 is closest to α_N and limits to α_N (see Section 4 and Appendix A). Furthermore, $\alpha_\ell < \alpha_N$ for all $\ell > 1$. Also recall that the Bonferroni inequality states that the family-wise error rate is less than the sum of the individual error rates: then, α_F can be no greater than the sum of the α_ℓ times the number of comparisons with true difference in the means equal to ℓ . That is,

$$\alpha_F \leq \sum_{\ell=1}^{k-1} \alpha_\ell (k - \ell). \tag{6}$$

The more conservative bound (5) is obtained by the fact that $\alpha_\ell < \alpha_1 < \alpha_N$. Since bounding α_F by adding the individual error rates is already a conservative procedure, and the α_ℓ decrease exponentially as ℓ increases, using the bound (5) is excessively conservative.

Creating a more accurate Bonferroni bound is not conceptually difficult, but it lacks the typical computational ease of the naive Bonferroni correction (5). Specifically, an upper bound on each α_ℓ, say $\tilde{\alpha}_\ell$, can be obtained numerically (for fixed values of n, k and α_N) by maximizing over the effect size, θ/σ, where σ is the common group-specific standard deviation. Then, the equation:

$$\alpha_D = \sum_{\ell=1}^{k-1} \tilde{\alpha}_\ell (k - \ell) \tag{7}$$

could be solved by modifying α_N to obtain the desired α_D, such as by a bisection algorithm.

Evaluations of the family wise error rates for a variety of effect sizes (described below) illustrates that the first term, $\alpha_1(k-1)$, is close to $\alpha_N(k-1)$ and completely dominates the right hand side of equation (6). Hence, a convenient and simple rule of thumb is to set α_N to $\alpha_D/(k-1)$. We refer to this multiple comparisons procedure as an ℓ-correction. Thus we contend that the naive Bonferroni procedure unnecessarily divides α_D by a factor of $k/2$. Below, we evaluate this rule of thumb and demonstrate that is much less conservative than a naive Bonferroni correction and is nearly equivalent to a correction based on (6).

4 Numerical evaluations

As is argued in Appendix A, we first note that α_1 limits to α_N as the effect size increases. Figure 1 displays the behavior of α_1 as a function of the parameter $\nabla = 2\theta/\left(\sigma\sqrt{2/n}\right)$. We chose this parameter rather than the effect size to match the notation of Schuirmann (1987). This figure illustrates that α_1 tends to the nominal error rate. Because of the square-root n in the denominator of the denominator of ∇, α_1 will typically be near α_N. Note that in this figure, and the remaining, the smallest possible degrees of freedom under our assumptions $(2n-2)$ was used.

Figure 2 displays the rapid decrease in error rate for those tests whose mean difference is $\Delta\mu = 2\theta$. Note that the maximum magnitude of these terms, $\tilde{\alpha}_2$, is on the order of 10^{-4}. Plots for larger values of ℓ are not shown, as their shape is similar with a rapidly decreasing maximum value. Figure 3 displays this decrease, by plotting $\tilde{\alpha}_\ell$ as ℓ increases.

Table 1 displays the bounds on the family-wise error rate for various values of n and k. This table shows the extreme conservatism of the naive Bonferroni bound (5), which is often well above 1. As it is also obtained by adding error rates, the bound based on (7) remains quite conservative, though is much less so than the naive Bonferroni bound. The next to the last column illustrates that the first term of (7) dominates the sum.

5 Example

We demonstrate the ℓ-correction on an example from the field of cell-engineering. Recently, scientists have been interested in comparing the effects of different labeling agents on what are called microcapsules (Barnett et al., 2007). Briefly, the function of a microcapsule is to deliver and house healthy xenogenic cells in patients whose own cells do not function properly, such as injecting porcine pancreatic cells into patients with type II diabetes. In order to monitor microcapsules once inside patients, labels that are either MRI (magnetic resonance imaging), ultrasound or X-ray visible are added to the micro-

capsules. However, researchers must asses the labels effect on the living cells inside the microcapsule. In one currently unpublished study, a human hepatic cell line (Hep G2 ATCC, Manassas) was encapsulated in contrast containing polyethylene glycol diacrylate microcapsules and their viability under 6 different labeling conditions was assessed. Included in this study was also an unlabeled control, making a total of 7 different conditions. The researchers were concerned with ensuring that the inclusion of contrast agent did not significantly alter viability of encapsulated cells and in assessing if cells were equally viable under the different labeling conditions. Hence, interest lies in testing biological equivalent viability between different labels in order to asses switch-ability. To test viability, cell survival was assessed at different time points after cell encapsulation and equivalence testing performed. Setting θ to 5%, all pairwise TOST test were performed, comparing all strata to each other at each time point. The result is a total of 21 comparisons per time point.

In Table 2 the larger of the two confidence endpoints, in absolute value, is given, both using the ℓ-correction and the naive Bonferroni correction. The TOST test could be performed for each by comparing these numbers to the tolerance limit. The 5% tolerance limit bound is above this upper value on several comparisons. Most importantly, the result of the test reverses after use of the less conservative multiplicity control in the comparisons: $(1, 2)$, $(5, 2)$, $(7, 2)$, and $(6, 4)$

6 Conclusion

The proposed ℓ-correction, simply setting the nominal error rate used in TOST to the desired family-wise error rate divided by one minus the number of groups compared, provides a fast and simple rule of thumb for testing the bioequivalence of multiple strata. The basis for this approach comes from a bound on the family-wise error rate by adding the individual error rates and noting that under a joint null-hypothesis for comparisons, only the $k - 1$ comparisons with the closest mean differences make any real contribution to

this bound. On a practical side, it is important to note that in the case where all strata are compared to a single control, the ℓ-correction and naive Bonferroni correction will be identical. However, in examples where all pairwise comparisons are made, such as the one considered above, the ℓ-correction will achieve a much tighter bound to the family-wise error rate.

We emphasize that, while a vast improvement over a naive Bonferroni correction, the proposed ℓ-correction is motivated by adding error rates and hence can be very conservative. Its main attractions are its ease of explanation and simple implementation. If these rationals are not of interest in the problem in hand, more optimal procedures should be pursued.

Acknowledgements We would like to thank Jean-Francois H. Geschwind MD and Bradley Barnett for the data set used in the example.

References

Anderson, S. and Hauck, W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics-Theory and Methods*, 12(23):2663–2692.

Barker, L., Luman, E., McCauley, M., and Chu, S. (2002). Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156(11):1056.

Barnett, B., Arepally, A., Karmarkar, P., Qian, D., Gilson, W., Walczak, P., Howland, V., Lawler, L., Lauzon, C., Stuber, M., et al. (2007). Magnetic resonance–guided, real-time targeted delivery and imaging of magnetocapsules immunoprotecting pancreatic islet cells. *Nature Medicine*, 13:986–991.

Barnett, B., Kraitchman, D., Lauzon, C., Magee, C., Walczak, P., Gilson, W., Arepally, A., and Bulte, J. (2006). Radiopaque alginate microcapsules for X-ray visualization and immunoprotection of cellular therapeutics. *Molecular Pharmaceutics*, 3(5):531–8.

Berger, R. and Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319.

Bofinger, E. (1985). Multiple comparisons and type iii errors. *Journal of the American Statistical Association*, 80(390):433–437.

Brown, L., Casella, G., and Hwang, J. (1995). Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Assocication*, 90:880–889.

Choi, L., Caffo, B., and Rohde, C. (2007). A survey of the likelihood approach to bioequivalence trials. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, page 134.

Fluehler, H., Grieve, A., Mandallaz, D., Mau, J., and Moser, H. (1983). Bayesian approach to bioequivalence assessment: an example. *Journal of Pharmaceutical Science*, 72(10):1178–81.

Giani, G. and Strassburger, K. (1994). Testing and selecting for equivalence with respect to a control. *Journal of the American Statistical Assocication*, 89:320–329.

Giani, G. and Strassburger, K. (2000). Multiple comparison procedures for optimally discriminating between good, equivalent, and bad treatments with respect to a control. *Journal of Statistical Planning and Inference*, 83(2):413–440.

Goodman, S. and Berlin, J. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121(3):200.

Hauck, W. and Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Pharmacodynamics*, 12(1):83–91.

Hoenig, J. and Heisey, D. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC.

Hsu, J., Hwang, J., Liu, H., and Ruberg, S. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, 81(1):103–114.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press Cambridge, UK.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Pharmacodynamics*, 15(6):657–680.

Selwyn, M., Dempster, A., and Hall, N. (1981). A Bayesian approach to bioequivalence for the 2 x 2 changeover design. *Biometrics*, 37(1):11–21.

Selwyn, M. and Hall, N. (1984). On bayesian methods for bioequivalence. *Biometrics*, 40(4):1103–1108.

Tempelman, R. (2004). Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies 1. *Journal of Animal Science*, 82(90130):162–172.

Westlake, W. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 32(4):741–744.

A Derivation of α_ℓ

We presume that the true value of $\Delta\mu$ is $\ell\theta$ and, for brevity, we denote the critical value $t_{df,1-\alpha_N}$ simply by t.

$$\begin{aligned}
\alpha_\ell &= P\left(\frac{\theta-\Delta\bar{X}}{s\sqrt{2/n}} > t \quad \text{and} \quad \frac{\Delta\bar{X}+\theta}{s\sqrt{2/n}} > t\right) \\
&= P\left(\frac{\theta(1-\ell)}{\sigma\sqrt{2/n}} - \frac{t}{\sqrt{df}}\sqrt{\frac{s^2df}{\sigma^2}} > \frac{\Delta\bar{X}-\ell\theta}{\sigma\sqrt{2/n}} \quad \text{and} \quad -\frac{\theta(1+\ell)}{\sigma\sqrt{2/n}} + \frac{t}{\sqrt{df}}\sqrt{\frac{s^2df}{\sigma^2}} < \frac{\Delta\bar{X}-\ell\theta}{\sigma\sqrt{2/n}}\right) \\
&= P\left(\frac{\theta(1-\ell)}{\sigma\sqrt{2/n}} - \frac{t}{\sqrt{df}}\chi^2_{df} > Z \quad \text{and} \quad -\frac{\theta(1+\ell)}{\sigma\sqrt{2/n}} + \frac{t}{\sqrt{df}}\chi^2_{df} < Z\right) \\
&= E\left[P\left(\frac{\theta(1-\ell)}{\sigma\sqrt{2/n}} - \frac{t}{\sqrt{df}}\chi^2_{df} > Z \quad \text{and} \quad -\frac{\theta(1+\ell)}{\sigma\sqrt{2/n}} + \frac{t}{\sqrt{df}}\chi^2_{df} < Z \;\middle|\; \chi^2_{df}\right)\right]
\end{aligned}$$

where, recall, df refers to the degrees of freedom and χ^2_{df} and Z represent independent chi-squared and Z random variables, respectively. A simple calculation yields that the interior probability is greater than zero only when $\frac{\theta^2 df}{t^2\sigma^2 2/n} > \chi^2_{df}$. Hence this may be written as:

$$\alpha_\ell = E\left[\left\{\Phi\left(\frac{\theta(1-\ell)}{\sigma\sqrt{2/n}} - \frac{t}{\sqrt{df}}\chi^2_{df}\right) - \Phi\left(-\frac{\theta(1+\ell)}{\sigma\sqrt{2/n}} + \frac{t}{\sqrt{df}}\chi^2_{df}\right)\right\} I\left(\frac{\theta^2 df}{t^2\sigma^2 2/n} > \chi^2_{df}\right)\right],$$

where $I(\cdot)$ is an indicator function. This formula was used for all calculations, with Gauss-Laguerre integration (see Press et al., 1992), implemented in R (Ihaka and Gentleman, 1996), used for the outer expectation.

Several points are in order: $i)$ because of the indicator function, this is not the difference between two non-central T probabilities; $ii)$ the only unknown that this equation depends on is the effect size θ/σ; $iii)$ the larger ℓ is, the smaller the two interior normal distribution probabilities are. $iv)$ small values of θ/σ will restrict the area of integration, hence yielding small probabilities; $v)$ for $\ell > 1$, large values of θ/σ will yield small probabilities for the interior normal distributions. Hence for $\ell > 1$, this probability will typically be very small, and decays exponentially fast as ℓ increases. Finally for $\ell = 1$, as θ/σ increases, the left normal term limits to α_N while the right one limits to 0.

B Tables

	Bound based on (7) for given value of n							Naive
k	5	10	15	20	50	1000	$(k-1)\alpha_n$	Bonferroni
6	.2513	.2509	.2509	.2508	.2508	.2508	0.25	0.75
12	.5524	.5521	.5520	.5520	.5520	.5520	0.55	3.30
20	.9538	.9536	.9536	.9536	.9536	.9536	0.95	9.50

Table 1: Values for bounds on the family-wise error rate when setting $\alpha_N = .05$ for various values of n and k. The first columns give the bound based on (7) while the second to the last column only uses the first term, $(k-1)\alpha_N$ and the naive Bonferroni correction is based on (5).

	1	2	3	4	5	6	7
1	-	5.49	7.39	6.34	4.44	6.36	4.95
2	4.52	-	7.92	7.43	5.43	7.54	5.77
3	6.44	6.86	-	9.34	7.33	9.44	7.69
4	5.55	6.43	8.35	-	6.49	5.01	6.93
5	3.68	4.45	6.38	5.65	-	6.52	4.89
6	5.58	6.57	8.49	4.16	5.72	-	7.01
7	4.07	4.75	6.68	6.01	4.01	6.12	-

Table 2: The maximum of the absolute value of the confidence interval for $\Delta\mu$ for the example 5. Hence a TOST test can be performed by comparing each number to the tolerance limit. Here the limits using the ℓ-correction are given below the diagonal while the limits using the naive Bonferroni are given above the diagonal. For example, the $[3, 1]$ and $[1, 3]$ cells give the comparison of groups one and three for the ℓ-correction and naive Bonferroni, respectively.

C Figures

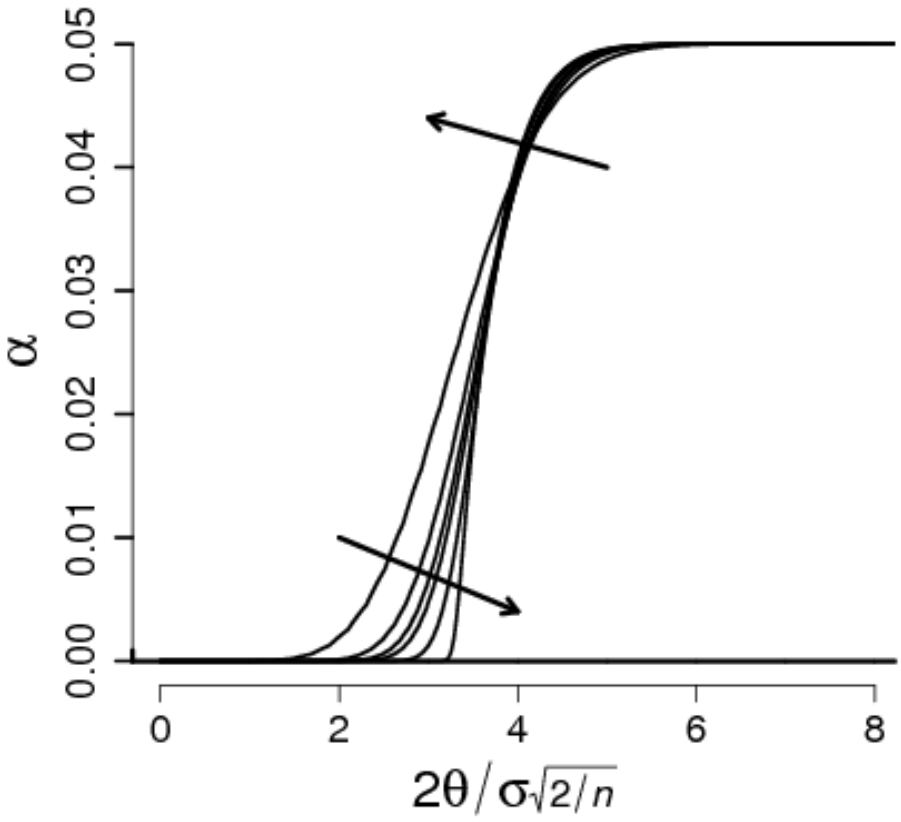

Figure 1: The true error rate, α_1, for a TOST test performed when $\Delta\mu = 1\theta$ plotted as a function of $2\theta/\sigma\sqrt{2/n}$. Each line represents a different sample size with $n = 5, 10, 15, 20, 50, 1000$. The arrows point in the direction of increasing n.

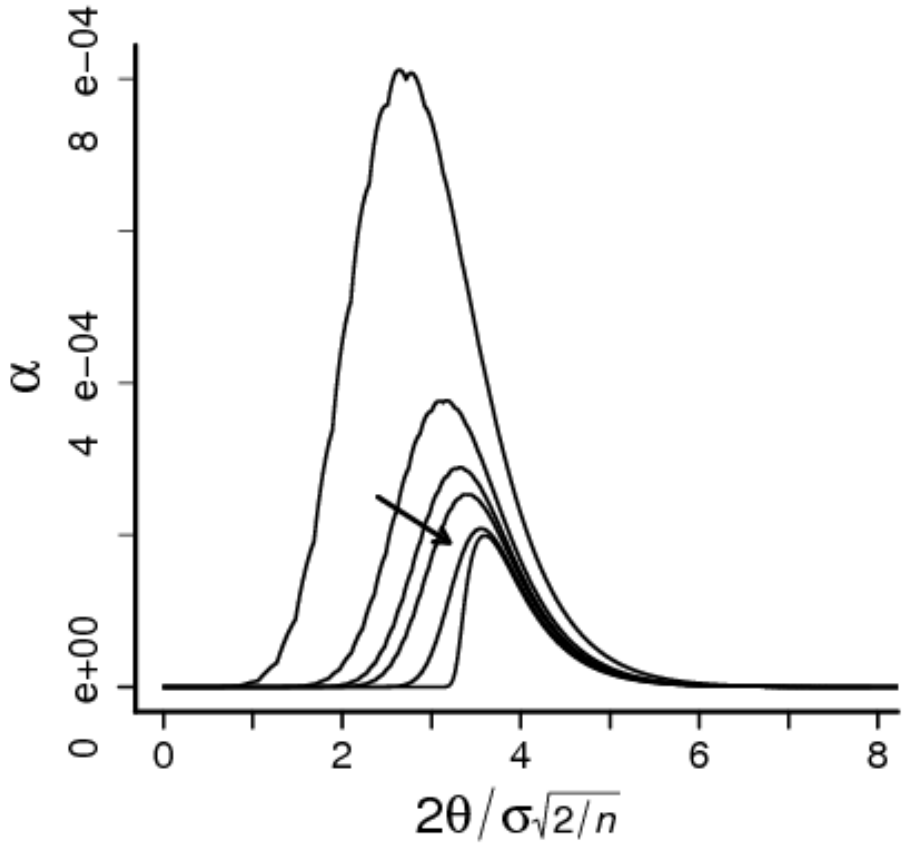

Figure 2: The true error rate, α_2, for a TOST test performed when $\Delta\mu = 2\theta$ plotted as a function of $2\theta/\sigma\sqrt{2/n}$. Each line represents a different sample size with $n = 5, 10, 15, 20, 50, 1000$. The shape for the $\ell = 2$ case is representative of the shapes of all plots for any case $\ell > 1$. The arrow points in the direction of increasing n.

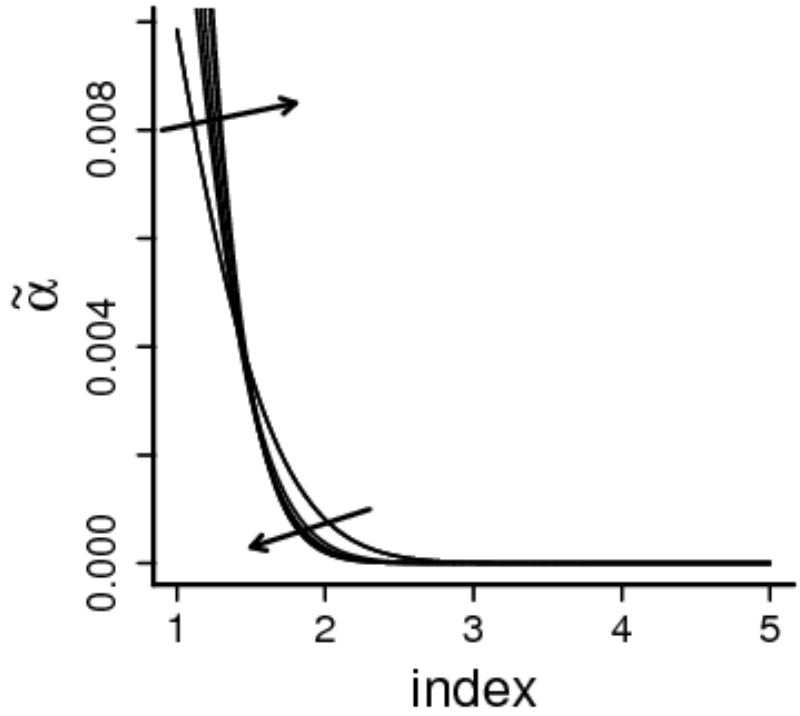

Figure 3: The maximum true alpha rates, $\tilde{\alpha}_\ell$, for a TOST test performed between two distributions whose means are $\ell\theta$ apart are shown for the six different sample sizes $n = 5, 10, 15, 20, 50, 1000$. Here ℓ is labeled "index" on the horizontal axis. A tolerance limit of five decimal places was used in the calculation of the maximums. The arrows point in the direction of increasing n.