Bioequivalence Testing

Thomas Oakberg

Statistics Writing Project

May 8, 2003

Before a generic drug or a new formulation of an existing drug can be placed on the market, it is necessary to show that the new drug is equally effective as the existing drug that is already on the market. The statistical test procedures used to determine if two formulations are equally effective are called bioequivalence tests. Two formulations are said to be bioequivalent if they are absorbed into the bloodstream at the same rate and if the concentration of the drugs in the bloodstream is the same for each of the two drugs.

Clearly, it is very important to show that a new drug is bioequivalent to the existing drug before placing the new drug on the market. The potential for serious dangers to the public exists if a generic or new drug is claimed to be bioequivalent to an existing drug when in fact it is not. It is very important that we have statistically sound bioequivalence test procedures, as well as appropriate confidence interval procedures associated with these tests.

Bioequivalence tests are generally performed by recording the concentrations of drugs in the bloodstream at set times after administration of the drugs. A curve can be generated for each drug displaying the concentration of the drug in the bloodstream vs. time after administration of the drug. We can then use this data to test to see if the two drugs are bioequivalent. Three variables are generally calculated from the data to be used in conducting a bioequivalence test. Using the notation from Berger and Hsu (1996), the first of these variables is AUC, which measures the area under the curve that describes the relationship between concentration and time. The other two variables are C_{max}, which is the maximum concentration of the drug in the bloodstream, and T_{max}, which is the time required to reach the maximum concentration occurs in the bloodstream. In order for two drugs to be declared bioequivalent, we would like these three variables to be similar for

the two drugs.

The hypotheses that are used in a bioequivalence test are quite different from the hypotheses that are generally used in most statistical tests, where the null hypothesis is a statement of "no effect", or "no difference". In a bioequivalence test, the null hypothesis states that there is a difference between the two types of drugs, and the alternative hypothesis is a statement of no difference between the drugs. There are multiple types of hypotheses that can be used in performing a bioequivalence test. The simplest type of hypotheses are called average bioequivalence hypotheses. Average bioequivalence hypotheses state that population means for the variables discussed above should be similar. Letting μ_T denote the mean AUC for the generic or new drug and μ_R denote the mean AUC for the existing drug, the standard hypotheses are:

$$H_0: \mu_T/\mu_R \leq \delta_L \text{ or } \mu_T/\mu_R \geq \delta_U$$

$$H_a: \delta_L < \mu_T/\mu_R < \delta_U$$

The values δ_L and δ_U are chosen so that the means must be "close enough" in order to declare bioequivalence. In the United States, the FDA has set the values of δ_L and δ_U at 0.80 and 1.25 respectively for both AUC and C_{max}. Europe uses the same values for AUC, but uses 0.70 and 1.43 for C_{max}. Often times, instead of using the means, logarithms are taken to form new hypotheses. These hypotheses are stated in terms of the difference between the two log means, as opposed to the ratio of the two true means. When using logarithms, the standard hypotheses for average bioequivalence are:

$$H_0: \eta_T - \eta_R \leq \theta_L \text{ or } \eta_T - \eta_R \geq \theta_U$$

$$H_a: \theta_L < \eta_T - \eta_R < \theta_U$$

In terms of these hypotheses, $\eta_T = \ln(\mu_T)$, $\eta_R = \ln(\mu_R)$, $\theta_L = \ln(\delta_L)$, and $\theta_U = \ln(\delta_U)$.

A second type of hypotheses that can be used in a bioequivalence test are called population bioequivalence hypotheses. Instead of just focusing on the population means for the two drugs, population bioequivalence hypotheses are statements about both the means and variances of the variables of interest. Letting σ^2_T denote the variance of AUC for the generic drug and σ^2_R denote the variance of AUC for the existing drug, these hypotheses can be expressed as:

$$H_0: \ \mu_T/\mu_R \leq \delta_L \text{ or } \mu_T/\mu_R \geq \delta_U$$

or

$$\sigma^2_T/\sigma^2_R \leq \kappa_L \text{ or } \sigma^2_T/\sigma^2_R \geq \kappa_U$$

$$H_a: \ \delta_L < \mu_T/\mu_R < \delta_U$$

and

$$\kappa_L < \sigma^2_T/\sigma^2_R < \kappa_U$$

In this case, both the means and variances must be similar in order to reject H_0. Since we are now involving variances, these hypotheses are more restrictive than the average bioequivalence hypotheses described above.

When performing a bioequivalence test, it is crucial that we can control the Type I error rate. In the context of bioequivalence, a Type I error would result in declaring two drugs to be bioequivalent, when in fact they are not. This would pose a much greater danger to the public than a Type II error, which would result in declaring two drugs to not be bioequivalent, when in fact they are bioequivalent. We must control the Type I error rate, α, at a specified level to limit the risk a Type I error would pose to the general public.

Many test procedures to test for bioequivalence have been devised, and each of which have their own advantages and drawbacks. Certain procedures are valid for different experimental designs, or methods in which the bioequivalence data is collected by the experimenters. While there are many different designs that can be used, the two simplest designs are parallel designs and two-period crossover designs. A parallel design occurs when two independent groups of subjects are given one of the two drugs. This is a very simple design, but the drawback is that there could be some difference between the two groups that could lead to incorrect results of the bioequivalence test. The two-period crossover design occurs when one group of subjects is administered the two drugs separately, and another group of subjects gets the two drugs in the reverse order. This design is more common than the parallel design.

What has become the standard test for average bioequivalence hypotheses was first proposed by Westlake and Schuirmann, and is called the "two one-sided tests", or TOST procedure. The main advantage of the TOST procedure is that it is fairly easy to perform. Letting D be an estimate of $\eta_T - \eta_R$ that is distributed normally with mean $\eta_T - \eta_R$ and variance σ^2_D, the test statistic for the TOST procedure is

$$t = \frac{D - (\eta_T - \eta_R)}{SE(D)}$$

where SE(D) is an estimate of σ_D that is independent of D and $r[SE(D)]^2/\sigma^2_D$ has a chi-squared distribution with r degrees of freedom. For a parallel design, $D = \overline{X} - \overline{Y}$, where $\overline{X}$ is the mean of the response measurements for the test drug and $\overline{Y}$ is the mean of the response measurements for the existing drug. The formulation of D is more complicated for a two-period crossover design, and depends on the model that is used for this design.

For the parallel design, the formula for the standard error of D is

$$se(D) = S\sqrt{\frac{1}{m} + \frac{1}{n}}$$

where m is the number of subjects in the test drug group, n is the number of subjects in the existing drug group, and S is the pooled estimate of the standard deviation. The standard error formula is the same for the crossover design, but the whole quantity is divided by two. It is important that the quantity D has the normal distribution for this test procedure, since the test statistic will involve quantities which have chi-squared and t-distributions. The process is resistant to slight departures from the normality assumption. In this case, the test statistic t has a Student's t distribution with r degrees of freedom. The TOST procedure performs two separate one-sided tests to determine bioequivalence, with two sets of hypotheses to be tested. The two tests are based on the statistics

$$T_U = \frac{D - \theta_U}{SE(D)} \quad \text{and} \quad T_L = \frac{D - \theta_L}{SE(D)}$$

which are computed using the same estimate D as defined above. The two sets of hypotheses to be tested are:

$$H_{O1}: \eta_T - \eta_R \leq \theta_L$$

$$H_{a1}: \eta_T - \eta_R > \theta_L$$

and

$$H_{O2}: \eta_T - \eta_R \geq \theta_U$$

$$H_{a2}: \eta_T - \eta_R < \theta_U$$

The null hypothesis for average bioequivalence is rejected if BOTH of the one-sided tests yield in a rejection of the null hypothesis in favor of the alternative. This happens if $T_U < -t_{\alpha,r}$ and $T_L > t_{\alpha,r}$, where $t_{\alpha,r}$ is the upper 100α percentile of the t-distribution with r

degrees of freedom.

Even though two separate level-α tests are performed when using the TOST procedure, it can be shown using the theory of Intersection-Union tests that the size of the overall test procedure is α. That is, the probability of a Type-I error is still α, and no adjustment needs to be done to the size of the two one-sided tests in order to control the Type-I error rate. The general set-up of an intersection-union test is that we wish to test the hypotheses

$$H_O: \quad \theta \in \bigcup_{i=1}^{k} \Theta_i$$

$$H_a: \quad \theta \in \bigcap_{i=1}^{k} \Theta_i^c$$

where $\Theta_1, \ldots, \Theta_k$ are subsets of the parameter space Θ. For each of the k subsets of the parameter space, we can construct a test of the hypotheses

$$H_{Oi}: \quad \theta \in \Theta_i$$

$$H_{ai}: \quad \theta \in \Theta_i^c$$

where each test of H_{Oi} against H_{ai} is an α-level test. Letting R_i denote the rejection region of the i^{th} test, it can be proven that the intersection-union test with rejection region $R = (R_1 \cap R_2 \cap \ldots R_k)$ is a level-α test of the intersection-union hypotheses. In terms of the TOST procedure, the individual rejection regions are $R_1 = (\theta_L, \infty)$ and $R_2 = (-\infty, \theta_U)$. The intersection of the two rejection regions is $R = (\theta_L, \theta_U)$, which is the rejection region for the intersection-union test, and therefore corresponds to a level-α rejection regions of the standard bioequivalence hypotheses.

With most test procedures, there exists a corresponding confidence interval procedure that can be used to get "similar" results to the test procedure. For example, for a standard two-sided one sample t-test, a 95% confidence interval for μ will contain the value in the null hypothesis exactly when a t-test using a .05 α-level will fail to reject the null. Similarly, the TOST procedure for bioequivalence has a corresponding confidence interval procedure that can be used. The standard confidence interval method corresponding to the TOST procedure, which has been supported by the FDA as the correct method to compute confidence intervals, is to use the interval

$$I = (D - t_{\alpha,r}SE(D), D + t_{\alpha,r}SE(D))$$

While this interval does correspond exactly to the rejection region of the TOST, it is a $100(1-2\alpha)\%$ interval, since $100(\alpha)\%$ of the distribution is being left off each end of the interval. We would like to come up with a procedure that would yield a $100(1-\alpha)$ procedure, since this would match up better with the size α TOST procedure.

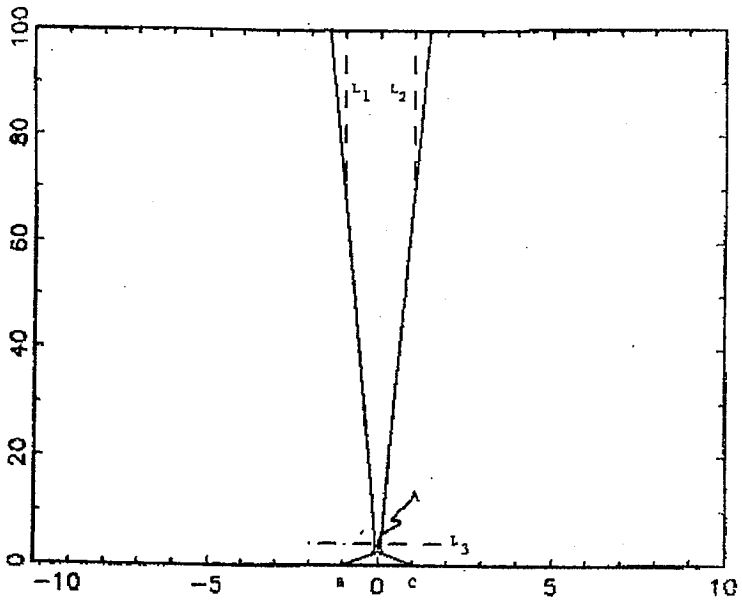
Hsu et. al (1994) and others have derived a modification of the $100(1-2\alpha)\%$ interval discussed above that results in a $100(1-\alpha)\%$ interval. This interval is given as

$$I^* = (\min(0, D - t_{\alpha,r}SE(D), \max(0, D - t_{\alpha,r}SE(D)))$$

If the interval I contains zero, then the interval I and the $100(1-\alpha)\%$ interval I^* will be the same. However, the two intervals will not be the same if the interval I does not contain zero. If the interval I lies to the right of zero, the interval I^* will extend from zero up to the upper endpoint of the interval I. If the interval I lies to the left of zero, the interval I^* will extend from the lower endpoint of I up to zero.

While the TOST procedure for determining bioequivalence has the advantage of being relatively simple to perform, it does have several serious disadvantages that have

Figure 1, rejection regions of the TOST and unbiased test



The rejection region of the TOST procedure is bounded by the triangle near the bottom of the graph. The rejection region of the unbiased test is bounded by all the solid lines on the graph. The rejection region of the TOST is completely contained within the rejection region of the unbiased test. Due to this fact about the rejection regions of the two tests, the unbiased test will be uniformly more powerful than the TOST. For small values of σ^2_D, the powers of the TOST and the unbiased test will be very close. However, as the value of σ^2_D increases, the difference between the power of the unbiased test and the power of the TOST can get quite large. Figures 2 and 3 show comparisons of the power functions of the unbiased test and the TOST. Figure 2 shows the power functions for a value of $\sigma^2_D = 0.4$, while Figure 3 shows the power functions for a value of $\sigma^2_D = 0.5$. The solid line represents the power function of the TOST, and the dashed line represents the power function of the unbiased test proposed by Brown, Hwang, and Munk.

led to attempts to find an "improved" procedure for testing for bioequivalence. The major drawback to the TOST procedure is that it has very small power for large values of σ^2_D. For large values of σ^2_D, the quantity SE(D) will be large, which will in turn lead to a small value of the test statistic. Since we need larger values of the test statistic in order to declare bioequivalence, the power of the test suffers greatly when σ^2_D is increased. Due to this lack of power, the TOST procedure is also biased, since an unbiased test should have the same power regardless of the value of σ^2_D. Table 1 gives values of the power of the TOST for $r = 30$ and $\alpha = 0.05$. Data taken from Berger and Hsu (1996)

Table 1

Value of σ_D	0.04	0.08	0.12	0.16	0.20
Power of TOST ($\eta_T - \eta_R = 0$)	1.000	0.720	0.158	0.007	0.000
Power of TOST ($\eta_T - \eta_R = 1.25$)	0.05	0.05	0.031	0.003	0.000

We can see that the TOST has essentially no power for larger values of σ_D. Clearly, this is not ideal. Several alternative test procedures have been proposed that are much more complex than the TOST, but do not suffer from the same lack of power as the TOST. Figure 1 shows the rejection regions for the TOST and an unbiased test proposed by Brown, Hwang, and Munk (1997). The estimate D of $\eta_T - \eta_R$ is plotted on the horizontal axis, and the estimated standard deviation of the estimate is plotted on the vertical axis. Figures 1, 2, and 3 are taken from Brown, Hwang, and Munk (1997).

Figure 2, comparison of power functions for $\sigma^2_D = 0.4$

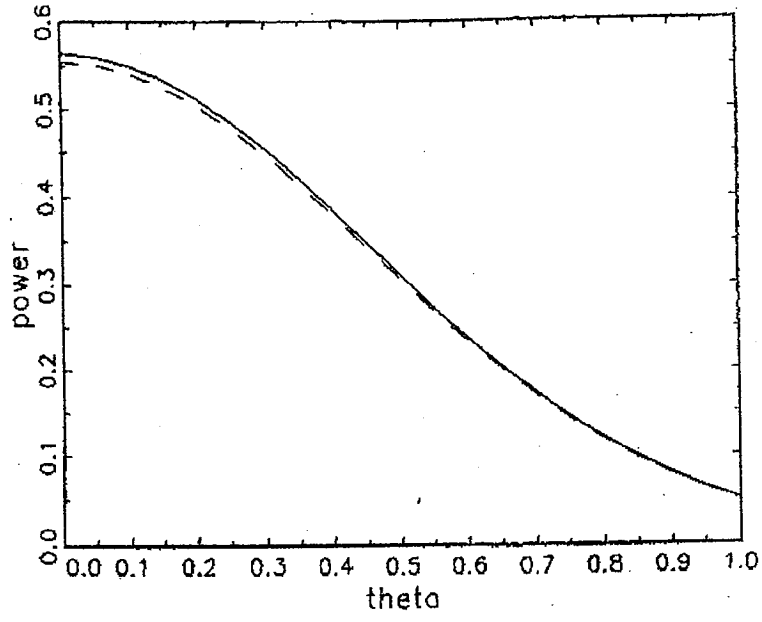
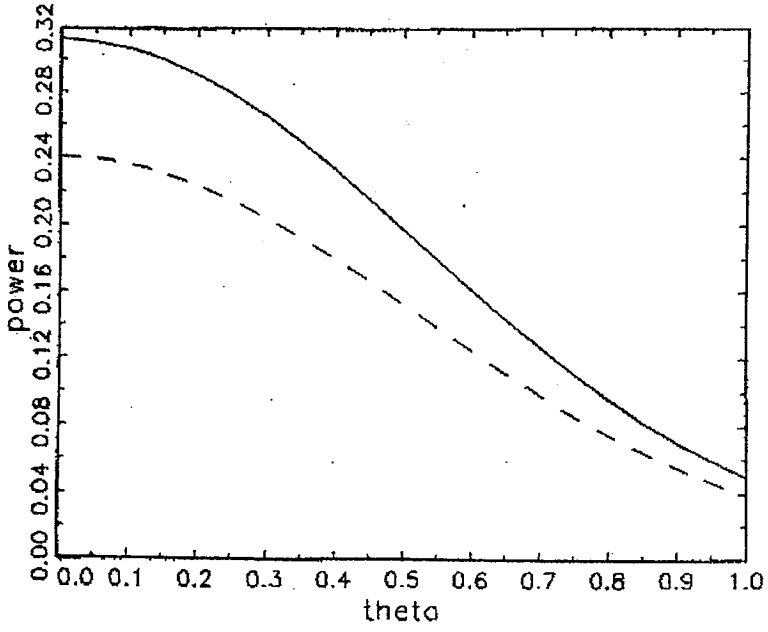


Figure 3, comparison of power functions for $\sigma^2_D = 0.5$



There have been many more test and confidence intervals proposed that the ones I have discussed. Each of these procedures has its advantages and disadvantages. As is frequently the case, the computational complexity of the test procedures seem to increase greatly as the properties of the test improve. The TOST is easy to perform but has some drawbacks, while the unbiased test proposed by Brown, Hwang, and Munk improves on the TOST but is much more difficult to perform. This increase in complexity makes it difficult to determine which of the test or confidence interval procedures is the "best" to determine bioequivalence. Due to the serious health risks that can be associated with an incorrect result of a bioequivalent test, the question of which procedure is the best one is an important one that needs to continue to be addressed.

References

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. *Statistical Science* **11**, 283-319

Brown, L.D., Hwang, J. T. G. and Munk, A (1997). An Unbiased Test for the Bioequivalence Problem. *The Annals of Statistics* **25**, 2345-2367

Hsu, J.C., Hwang, J. T. G., Liu, H.K. and Ruberg, S. J. (1994). Confidence Intervals associated with tests for bioequivalence. *Biometrika* **81** 103-114