Demonstrating Statistical Equivalence of a Rapid Microbiology Method to a Compendial Method

David A. Porter, Ph.D.

There are many possible uses for rapid microbiology methods (RMM). There are also many potential purposes behind incorporating such methods into routine operations of a pharmaceutical microbiology laboratory. The intention for the use of an RMM might be to provide for faster screening of actives/formulations. Perhaps the purpose might be to obtain more data of higher quality at a faster rate than the conventional method. It is also possible that the intention is to use the RMM as a replacement for a compendial method. The goal of this article is to suggest a statistical approach for demonstrating equivalence of an RMM to its respective compendial method.

The requirement for the use of an alternative method (RMM in this case) to a compendial method is that the alternative method must be shown to be at least equivalent to the compendial method (*General Notices and Requirements*, *USP 30*[1]). Many people set out to demonstrate statistical equivalence of an RMM to a compendial method by performing head-to-head studies, then examining the results for the existence of statistical differences between the results. If no such results are observed, the conclusion reached is that because there are no statistically significant differences between the results, the methods are statistically equivalent. It turns out that using such a statistical approach is not correct.

As you may recall from introductory statistics classes, statistical testing for differences between two sets of data first requires that a null hypothesis and an alternative hypothesis be generated. The approach you most likely learned in introductory statistics (for testing for a significant difference between averages) is to make the null hypothesis state "There is no difference between the means", and the alternative hypothesis "There is a significant difference between the means". With the two hypotheses set up in this manner, if a significant difference is found between the means, you reject the null hypothesis and accept the alternative hypothesis. However, if you do not find evidence of a statistically significant difference between the means, you should not state that you "accept" the null hypothesis. Instead, you should state that you cannot "reject" the null hypothesis. This is very much like the situation in the American judicial system, where a verdict of "not guilty" is not a statement that the defendant was found innocent. Statistically, one can prove the alternative hypothesis, or fail to reject the null hypothesis.

What all of this means in the case of proving equivalence is that it is necessary to essentially reverse the null and alternative hypotheses. Thus, the null hypothesis becomes "There is a significant difference between the means", and the alternative hypothesis becomes "There is no significant difference between the means". This makes it possible to prove the alternative hypothesis (the methods are not significantly

different), or fail to reject the null hypothesis (there may or may not be significant differences, you just haven't proven which is the case).

Let's make certain that the notion of "equivalence" is well spelled out here. By stating that two methods are "equivalent", we are not stating that the results from the methods are identical. Rather, we are stating that the methods are sufficiently similar[2]. This is an extremely important point to grasp because it lies at the core of statistical proof of equivalence. As microbiologists, you know that the conventional microbiological methods have lots of variance (noise) associated with them, much of which is not scientifically meaningful. For example, using the plate count method it is unlikely that there is a real difference between 50 and 60 counts. In fact, the newly harmonized chapter <61> in *USP*[3] contains a section in which it is stated that, for example, if the acceptance criterion for an article (as found in a monograph) is 100 cfu, the maximum acceptable count is 200 cfu. Therefore, what must be established scientifically is how similar is similar enough. By making this determination you are essentially establishing "goal posts". If the results obtained with the RMM fall within the goal posts, the methods are determined to be "equivalent".

In many cases, determining the location of the goal posts is the most difficult part of equivalence testing. However, in some cases when comparing a compendial method with an RMM, it might not be so difficult. Take, for example, an RMM intended to substitute for the plate count method as per chapter <61>. In this case, you have in the text a section (alluded to above) suggesting that a range of ± 0.3 log is acceptable. By choosing 0.3 log to set up the goal posts, you are stating that a difference of ± 0.3 does not constitute a scientifically significant difference. This represents the allowable difference, or as is often referred to, the Δ.

Next, let's look at how we could perform the statistical analysis for equivalence. Two approaches will be described, one using two one-sided t-tests (TOST), and the other using 95% confidence intervals.

The form of the t-test you most likely have seen before is:

T = difference between means/standard error of difference

The form of the t-test to be used in equivalence testing is:

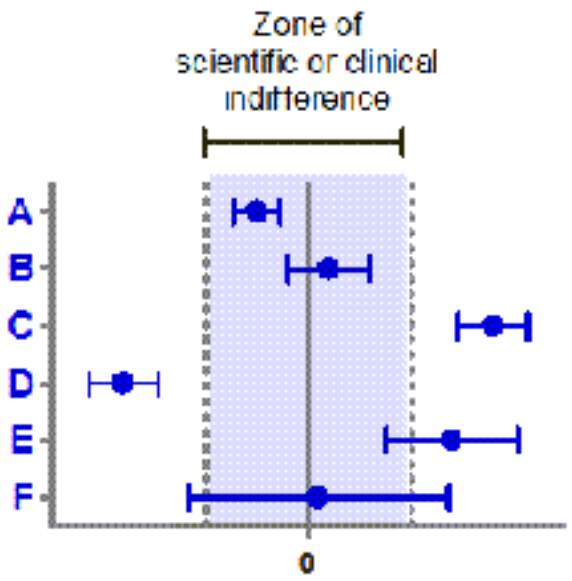
T = (difference between means) + Δ /standard error of difference)

and

T = (difference between means) - Δ /standard error of difference

Note that the critical difference between the two forms of t-test is the inclusion of the Δ, the amount of difference that is tolerable, in the second form.

The use of the 95% confidence interval approach is one I prefer because it is more pictorial. Examine the figure below, copied from a web site[4]:



For our purposes, imagine that the two dashed vertical lines (goal posts) represent ± 0.3 log around the mean of counts obtained by the compendial plate count method. The results from six different methods, labeled A-F, are depicted as 95% confidence intervals. If the 95% confidence intervals for the alternative method fall entirely within the goal posts, then you accept the alternative hypothesis that there is no statistically significant difference between the two methods. By basing the Δ on your previous knowledge of what constitutes a scientifically meaningful difference using the compendial method (or by using information provided in chapter <61>), you have already accounted for differences that may appear in the numbers that may represent only noise.

What can you say about the imaginary results from the six methods depicted above? The 95% confidence interval (CI) for method A falls entirely within the goal posts, therefore you would state that the alternative hypothesis is accepted (no meaningful difference between the alternative and compendial methods). Note here that had you just done a t-test using the traditional arrangement of hypotheses, a conclusion might have been reached that the methods were significantly different given that the 95% CI of method A did not cross the mean for the compendial method. However, by first determining what difference is scientifically meaningful, you can see that this is indeed a case of statistical significance not being indicative of a scientifically meaningful difference.

The 95% CI results for method B fall clearly within the goal posts, therefore the compendial and alternative method would be declared to be equivalent. As before, if you were to compare methods A and B using the traditional approach with the t-test, you might conclude that these two methods differ. But again, the results for both methods fall within the goal posts, so relative to the compendial method, neither method differs from the compendial method.

The 95% CI results for method C fall entirely to the right of the goal posts. You would therefore not reject the null hypothesis. Remember that by doing so, you are not proving that there is a significant difference between the alternative and compendial methods, just that you cannot prove that there is no difference. Looking at the array of results for method C all being to the right of the goal posts, you might decide to set up an experiment where you make an alternative hypothesis stating that the alternative method provides significantly higher counts than the compendial method, and a null hypothesis of there being no significant difference. In this case, you would revert to the traditional approach.

The 95% CI results for method D fall entirely to the left of the goal posts. Your conclusions would be similar to those for method C, except that were you to continue with this method, the new alternative hypothesis would be that the alternative method provides lower counts. Why you would want to show this is beyond me!

The results for methods E and F are ambiguous in that part of the 95% CIs are within the goal posts, part outside. You might want to conduct further experiments with such data. Method F does appear to be less precise based upon the width of its CI, but again, you will not have proven that with this experiment.

The take home message is this. If you have reason to believe the RMM is superior to the compendial method, by all means set up your hypotheses in the traditional manner. If instead your interest is in proving equivalence, then remember to, in essence, reverse the hypotheses. Then determine the range of results within which if the 95% confidence interval for you RMM falls, you will have proven statistical equivalence. However, remember that as is always the case with statistics, there is a possibility that you will reach the wrong conclusion. That is always a possibility given that statistical analysis is based upon limited sample sizes, and therefore by chance you may obtain data that would fall within the goal posts. Do not use statistics in place of good scientific judgment.

References

[1] *ThePharmacopeia of the United States of America, 30th Revision*

[2]Presentation by Walter W. Hauck, Ph.D. at the USP Conference on Biological and Biotechnological Drug Substances and Products, November 21, 2003

[3]*Microbiological Examination of Nonsterile Products: Microbial Enumeration Tests, USP 30*

[4]http://www.graphpad.com/library/BiostatsSpecial/article_182.htm