Controlling the Alpha Risk of Multiple Equivalence Tests Using the Intersection-Union Test (IUT) Principle

By Joel Dobson and Thomas D. Murphy

Q When making simultaneous multiple statistical hypothesis tests, the *α* risk levels for each of the *n* component tests are often adjusted by the Bonferroni Correction in order to meet an overall *α* risk for the family of the combined tests. This is accomplished by dividing each component test *α* risk by *n*. For example, if five outlier tests are conducted at an overall risk of 0.05, then each component test is conducted at α = 0.05/5 = 0.01. How does equivalence testing avoid this multiplicity correction by utilizing the IUT principle?

A The IUT principle and its application to multiple testing requirements has been given by Berger[1] to alleviate this multiplicity problem, in particular for simultaneous multiple equivalence tests. The solution is to reverse the null hypothesis for each of the component tests such that its requirement is not met at a given α risk, then rejection of every null hypothesis accepts that all requirements are met. The overall null hypothesis test posits non-equivalence for the combined tests and its alternative accepts equivalence. The overall test rejects its null hypothesis of overall non-equivalence, and decides that all of the requirements are met, if and only if each of the individual tests decide that its requirement is met. To see this, let R_i (i = 1, ..., *n*) be the event that null hypothesis of the i-th statistical test is rejected. Then the event $R = \bigcap_{i=1}^{n} R_i$ for the overall test is the *intersection* of all R_i events, which is the event that all *n* null hypotheses are rejected at risk *α*, thus accepting that all requirements are met. Note that the complementary event $R^c = \bigcup_{i=1}^{k} R_i^c$ is the *union* of all complementary R_i^c events that their null hypotheses are not rejected at risk *α*. The R^c event defines the null hypothesis of the overall test, and is accepted when at least one or more individual null hypotheses are accepted, thus rejecting the meeting of all requirements. The overall test is an IUT with the null hypothesis as the union and the alternative hypothesis as the intersection. If the component tests are all one-sided, the overall test and all component tests will have risk *α*.

Example: Standard practice for evaluating equivalence of two testing processes (E2935) uses numerical data from two sources of test results to determine if their true means, variances, or other parameters differ by no more than predetermined limits. As an example, the equivalence evaluation of mean differences in test results between two laboratories will be given. Define the true mean difference between labs as $\Delta = \mu_2 - \mu_1$, and set an equivalence limit E as the minimum tolerable difference. The null hypothesis for the overall statistical test is H_0: $|\Delta| \geq E$, that the true mean difference is equal to or greater than E, and the alternative hypothesis is H_a: $-E < \Delta < E$, that the true mean difference is less than ±E. Note that this is the reverse of the usual hypothesis setup for a statistical test for a zero difference in means.

The overall test consists of a two one-sided statistical test (TOST) procedure with the null and alternative hypotheses of the two component tests as follows in Table 1.

	Test 1	Test 2
Null hypotheses	H_{01}: $\mu_2 - \mu_1 \geq E$	H_{02}: $\mu_2 - \mu_1 \leq -E$
Alternative hypotheses	H_{a1}: $\mu_2 - \mu_1 < E$	H_{a2}: $\mu_2 - \mu_1 > -E$

Table 1

The TOST is an IUT because of the special way that the null hypotheses are defined for the component tests.

Test result data are collected from two testing Labs 1 and 2 and the averages from each lab are calculated. The difference *D* between the two averages and the standard error of the difference s_D are also calculated. The t statistics are $t_1=(E-D)/s_D$ and $t_2=(E+D)/s_D$ for Tests 1 and 2, respectively. Both null hypotheses are rejected when $t_1 > t$ and $t_2 > t$, where $t = t_{1-\alpha,f}$ is the upper $(1-\alpha)^{th}$ quantile of the Student's t distribution with *f* degrees of freedom. If both hypotheses are rejected, then it is asserted that $-E < \mu_1 - \mu_2 < E$ and the two sources are said to be equivalent; otherwise, the two data sources are deemed non-equivalent. Each hypothesis is tested at level *α*.

The TOST is also at level α because the rejection region (where equivalence is asserted) is the intersection of rejection regions for these two tests and therefore has probability ≤*α* under both of the null hypotheses. On the test result scale of the data the rejection regions are:

$$R_1 = (-\infty, E - t\,s_D),\ R_2 = (-E + t\,s_D, \infty), \text{ and}$$
$$R = R_1 \cap R_2 = (-E + t\,s_D, E - t\,s_D).$$

Any value of *D* within region *R* will accept means equivalence between Labs 1 and 2. It can be shown that the overall test level is exactly *α* under mild conditions.[2]

A previous Datapoints article showed means equivalence testing in terms of 100(1-2α)% confidence intervals instead of statistical hypothesis testing, which may be more intuitive because confidence intervals convey the uncertainty of the *D* estimate.[3] To meet means equivalence, the confidence interval, $D \pm t\,s_D$, must be completely contained within the interval (*-E, E*). Berger and Hsu criticize the use of confidence intervals in equivalence evaluation, but they do concede that the confidence interval approach for TOST is valid if the confidence interval is equal-tailed; that is, the same alpha is used for each half-interval.[4] The IUT principle can be extended to other applications of multiple comparisons and is usually more relevant than the statistical test for zero differences. ■

REFERENCES

[1] Berger, R., "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics,* Vol 24, No. 4, 1982, pp. 295–300.

[2] Ibid.

[3] Murphy, T.D., "Testing for Equivalence," *ASTM Standardization News,* Sept./Oct. 2014, Vol. 42, No. 5, pp. 16-17. 2.

[4] Berger, R., and Hsu, J., "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science,* Vol 11, No. 4, 1996, pp. 283–319.

Joel Dobson is a reliability engineer for Texas Instruments Incorporated, where he is a distinguished member of the technical staff. Dobson is an accredited professional statistician per the American Statistical Association, a certified six sigma green belt and black belt, a certified quality engineer per ASQ, and a member of the ASTM quality and statistics committee (E11).

Thomas D. Murphy is a retired statistical consultant and chairman of the subcommittee on test method evaluation and quality control (E11.20), a part of the committee on quality and statistics (E11). He served as chairman of E11 from 2002-2003 and is an ASTM Fellow**.**

John Carson, Ph.D., of P&J Carson Consulting LLC, is the Data Points column coordinator. He is chairman of the subcommittee on statistical quality control (E11.30), part of the committee on quality and statistics (E11), and also a member of the committee on environmental assessment, risk management, and corrective action (E50).