Paper 11683-2016

Equivalence Tests

Fei Wang and John Amrhein, McDougall Scientific Ltd.

ABSTRACT

Motivated by the frequent need for equivalence tests in clinical trials, this paper provides insights into tests for equivalence. We summarize and compare equivalence tests for different study designs, including designs for one-sample problem, designs for two-sample problem (paired observations, and two unrelated samples), and designs with multiple treatment arms. Power and sample size estimation are discussed. We also give examples to implement the methods using the FREQ, TTEST, MIXED, and POWER procedures in SAS/SAT® software.

INTRODUCTION

Equivalence testing is a natural approach to many statistical problems. Consider the following example: a drug has been in successful use for many years. A generic version that is less expensive is developed. To obtain approval of this generic version, a proof of similarity or non-existence of differences in efficacy between the two drugs is required.

The traditional hypothesis testing, "difference testing", is to demonstrate a difference between two groups or treatments. The null hypothesis states equality, versus an alternative hypothesis stating the existence of differences. For example, let μ_1 and μ_2 denote the mean of two treatments, respectively. The null hypothesis is:

$$H_0: \mu_1 = \mu_2;$$

and the two-sided alternative hypothesis is:

$$H_0: \mu_1 \neq \mu_2.$$

The traditional hypothesis testing is not an appropriate method for a study with the objective of demonstrating the non-existence of differences (equivalence). If the computed p-value is greater than a pre-specified significant level, the null hypothesis will not be rejected. However, non-rejection of the null hypothesis is not a sufficient proof of its validity, and non-rejection of the null hypothesis only indicates that there is not enough evidence to conclude the difference. For example, it could be that the sample size is too small to declare a significant inherent difference. In such a situation, post-hoc power calculations can be done to give more credibility of the null hypothesis. However, observed power calculations may not have any evidential value (Hoenig et al. 2001).

Equivalence testing is an approach which aims to demonstrate similarities or lack of differences. That is, instead of establishing a null hypothesis of "equality" versus an alternative hypothesis of "non-equality", equivalence testing establishes a null hypothesis of "sufficiently large difference" versus an alternative hypothesis of "near equality". The null hypothesis of an equivalence test is:

$$H_0: (\mu_2 - \mu_1) \leq -\varepsilon \; or \; (\mu_2 - \mu_1) \geq \varepsilon;$$

and the alternative hypothesis is:

$$H_a: -\varepsilon < (\mu_2 - \mu_1) < \varepsilon,$$

where ε $(\varepsilon > 0)$ is a pre-specified limit, and the equivalence margin $(-\varepsilon, \varepsilon)$ characterizes the range for the acceptable difference. Non-symmetric equivalence limits can be specified as $(\varepsilon_1, \varepsilon_2)$, where $\varepsilon_1 < \varepsilon_2$.

If the null hypothesis is rejected, the alternative hypothesis is accepted and the equivalence can be claimed.

The rest of this paper discusses popular methods of equivalence testing and their implementations in SAS.

DESIGNS FOR ONE-SAMPLE PROBLEM

Suppose you are interested in testing whether a binomial proportion is equivalent with a reference value (e.g. 0.65). The acceptable deviation from the reference value (equivalence margin) is 0.05.

Let p and p_0 denote the binomial proportion and the reference value, respectively. The hypotheses for the equivalence test of the binomial proportion are:

$$H_0: p \le p_0 - \varepsilon \; or \; p \ge p_0 + \epsilon;$$
$$H_a: p_0 - \varepsilon < p < p_0 + \varepsilon;$$

where ϵ is the pre-specified equivalence margin ($\varepsilon = 0.05$). You can use the Two One-sided Test (TOST) procedure to test for the equivalence of the proportion with the reference value. That is, a right-sided test for the lower equivalence margin with hypotheses:

$$H_{01}: p \le p_0 - \varepsilon;$$
$$H_{a1}: p > p_0 - \varepsilon;$$

and a left-sided test for the upper equivalence margin with hypotheses:

$$H_{02}: p \ge p_0 + \epsilon;$$
$$H_{a2}: p < p_0 + \varepsilon.$$

The asymptotic Wald test statistics, for the right-sided test and the left-sided test, are computed as:

$$Z_L = \frac{(\hat{p} - p_L)}{se(\hat{p})} \ge Z_\alpha \; and \; Z_U = \frac{(\hat{p} - p_U)}{se(\hat{p})} \le Z_\alpha$$

where $p_L = p_0 - \varepsilon$, $p_U = p_0 + \varepsilon$, $se(\hat{p})$ is the standard error which can be computed as $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$, n is the sample size, and α is the significance level.

The p-value for the equivalence test is the larger p-value from the above two one-sided tests, which is $P = \max(P_L, P_U)$, and $P_L = Prob(Z > Z_L)$, $P_U = Prob(Z < Z_U)$. If the p-value is smaller than α (which is typically 0.05), then non-equivalence hypothesis is rejected and an equivalence conclusion between the binomial proportion and the reference value can be claimed.

The equivalence can also be declared if the ordinary $100(1-2\alpha)\%$ confidence interval for the proportion is completely contained in the equivalence interval $(p_0 - \varepsilon, p_0 + \varepsilon)$.

In SAS, the FREQ procedure can perform the Two One-sided Test for the equivalence of proportion. To give an example, we first simulate a sample of 500 observations from a Bernoulli distribution with p=0.63. The SAS code to perform the equivalence test is:

```
proc freq data=example1;
  table yn / binomial(equiv p=0.65 margin=0.05);
run;
```

In the PROC FREQ step, YN is a binary variable with 1 (yes) and 2 (no), the option BINOMIAL(EQUIV P= MARGIN=) on the TABLE statement requests an equivalence test for the binomial proportion, with the reference value specified with the P=value option and the equivalence margin specified with the MARGIN=value option.

Output 1 shows the output from the above PROC FREQ step. The estimated proportion is 0.6440 with standard error 0.0214. Because the overall p-value for the equivalence test is 0.0199 (<0.05), the null hypothesis of non-equivalence is rejected. The 90% confidence interval (0.6088, 0.6792) is completely contained in the equivalence interval (0.6, 0.7). Therefore, the equivalence of the proportion with the reference value can be declared.

```
                    Equivalence Analysis
          H0: P - p0 <= Lower Margin or >= Upper Margin
            Ha: Lower Margin < P - p0 < Upper Margin

      p0 = 0.65   Lower Margin = -0.05   Upper Margin = 0.05


                    Proportion    ASE (Sample)
                       0.6440          0.0214

                  Two One-Sided Tests (TOST)
          Test                    Z         P-Value
          Lower Margin       2.0548    Pr > Z   0.0199
          Upper Margin      -2.6152    Pr < Z   0.0045
          Overall                              0.0199

           Equivalence Limits    90% Confidence Limits
            0.6000     0.7000        0.6088     0.6792

                       Sample Size = 500
```

Output 1. Output from a PROC FREQ Step

You can compute the power for this Two One-sided Equivalence Test using the POWER procedure. An example of the SAS code is:

```
proc power;
  onesamplefreq test=equiv_z method=normal
    lower  = 0.6
    upper  = 0.7
   proportion = 0.65
    varest = sample
    ntotal = 500
    power  = .
    alpha  = 0.05;
run;
```

The ONESAMPLEFREQ statement performs power and sample size analysis for equivalence tests by specifying TEST=equiv_z. Output 2 shows the output from the PROC Power step. The computed power is 0.516, which is very small. You may want to increase the sample size in order to have a test with larger power (see Figure 1).

```
          Z Test for Equivalence of Binomial Proportion

                    Fixed Scenario Elements
        Method                    Normal approximation
        Variance Estimate              Sample Variance
        Lower Equivalence Bound                   0.6
        Upper Equivalence Bound                   0.7
        Alpha                                    0.05
        Binomial Proportion                      0.65
        Total Sample Size                         500


                        Computed Power
                            Power
                            0.516
```

Output 2. Output from a PROC POWER Step

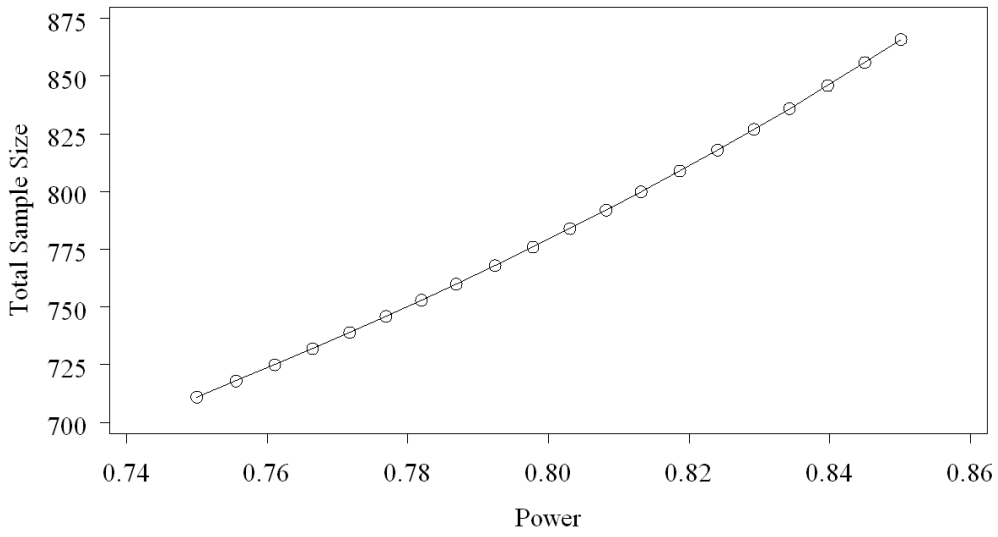

Figure 1. Equivalence Test for One-sample Test of Proportion

DESIGNS FOR TWO-SAMPLE PROBLEM

Consider the following example: in an experimental study of the effects of increased intracranial pressure on the cortical micro flow of rabbits, a preliminary test had to be done to ensure that the measurements exhibit sufficient stability during a pre-treatment period of 15 minutes' duration (Wellek, 2003).

Let X_i and Y_i denote the ith measurements of flows at the beginning and the end of that period, respectively, $i = 1, \cdots, n$. Let D_i denote the difference between the measurements. The analysis is similar to the analysis for the one parameter problem, by assuming that D_i follows a normal distribution and applying the TOST technique.

With the pre-defined equivalence margin ε (e.g. $\varepsilon = 0.2$), the null hypothesis of the equivalence test is:

$$H_0: \delta \leq -\varepsilon \; or \; \delta \geq \varepsilon;$$

and the alternative hypothesis is:

$$H_a: -\varepsilon < \delta < \varepsilon;$$

where δ is defined as the mean difference between the measurements at the end and beginning of the periods.

Suppose the flows of a sample of 25 animals are measured at the beginning and the end of that period. The observed mean flow are 53.01 and 52.96 at the beginning and the end of the period, respectively, corresponding to a mean change of 0.05. The standard deviation of the mean change is 0.27.

The equivalence test for a paired-sample comparison of means can be conducted using the PAIRED statement in the TTEST procedure:

```
proc ttest data=example2 dist=normal tost(-0.2,0.2);
  paired before*after;
run;
```

where DIST=NORMAL option on the PROC TTEST statement specifies a normal distribution for the data, TOST(-0.2, 0.2) requests a TOST equivalence test with the lower equivalence margin -0.2 and the upper equivalence margin 0.2.

Output 3 shows the output from the above PROC TTEST step:

```
                          The TTEST Procedure

                    Difference:  BEFORE - AFTER
         N        Mean     Std Dev     Std Err     Minimum     Maximum
        25      0.0537      0.2734      0.0547     -0.6891      0.4387

           Mean       95% CL Mean        Std Dev      95% CL Std Dev
         0.0537     -0.0591   0.1665      0.2734      0.2134   0.3803



                TOST Level 0.05 Equivalence Analysis

                Lower                          Upper
          Mean  Bound       90% CL Mean        Bound Assessment
        0.0537   -0.2 <  -0.0398   0.1472 <    0.2 Equivalent

             Test         Null        DF    t Value    P-Value
             Upper        -0.2        24       4.64     <.0001
             Lower         0.2        24      -2.68     0.0066
             Overall                                    0.0066
```

Output 3. Output from a PROC TTEST step

It can be seen from Output 3 that the 90% confidence interval (-0.0398, 0.1472) is completely contained within the equivalence interval (-0.2, 0.2). The overall p-value, 0.0066, is much smaller than 0.05. A conclusion of equivalence between the beginning and the end period in the measurements of flow can be declared.

You may want to estimate the sample size for the design with paired observations. The following is an example of SAS code which computes sample sizes for different mean differences and standard deviations and requests a plot of the sample size estimation:

```
proc power;
  pairedmeans test=equiv_diff
    lower    = -0.3
    upper    = 0.3
    meandiff = -0.2 to 0.2 by 0.05
    stddev   = 0.2 to 1.5 by 0.5
    corr     = 0.85
    npairs   = .
    alpha    = 0.05
    power    = 0.80
    ;
   Plot x=effect step=0.05
        Vary(color by stddev)
        Yopts=(ref=0 to 250 by 50);
run;
```

The TEST=EQUIV_DIFF option requests power analysis for an equivalence test based on the difference of means. The PLOT statement plots the number of pairs needed for each of the scenarios, see Figure 2. It shows that the larger the standard deviation or the farther the mean difference from 0, a larger sample size is required for the equivalence test.

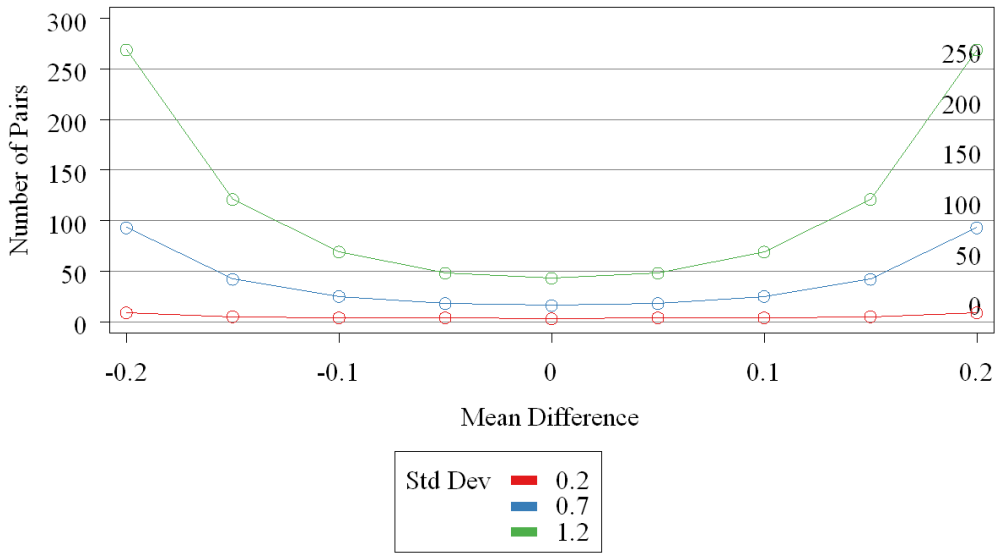

Figure 2. Equivalence Test for Paired Mean Difference

You can also perform equivalence tests for two unrelated samples, using the hypotheses discussed in the introduction section. The CLASS and VAR statements in PROC TTEST are required for the equivalence test for two unrelated samples.

DESIGNS WITH MULTIPLE TREATMENT ARMS

Consider the following example: an animal study conducted under the stability monitoring program is intended to evaluate the stability (equivalence) of the immune response induced by an influenza vaccine over a 12 month period. The immune responses of the vaccine, which was serially diluted at 6 dose levels, were measured at 0, 3, 6, and 12 months following the production with 2 replications. The aim of the study is to demonstrate the overall stability of the immune response.

The F-test for equivalence is an appropriate approach as a global test when comparing the multiple (i.e. > 2) means. The F-test relies on the assumption that the k groups of observations are normally distributed with means $\mu_i, (i = 1, \cdots, k)$, and common variance σ^2.

The null hypothesis states non-equivalence:

$$H_0: \varphi^2 \geq \varepsilon^2;$$

The alternative hypothesis is:

$$H_a: \varphi^2 < \varepsilon^2,$$

where: ε is the pre-specified equivalence margin ($\varepsilon > 0$), φ^2 is a weighted and standardized squared difference given by $\varphi^2 = \frac{\sum_{i=1}^{k}(n_i/\bar{n})(\mu_i - \tilde{\mu}_\cdot)^2}{\sigma^2}$, n_i is the number of observations in group i, $\bar{n} = \sum_{i=1}^{k} n_i/k$, and $\tilde{\mu}_\cdot = \sum_{i=1}^{k} n_i\mu_i / \sum_{i=1}^{k} n_i$.

Psi-squared, φ^2, is a suitable standardized (normalized) measure of difference between group/replicate/time point means and an overall (across all groups/replicates/time points) mean. Wellek (2003) develops the test using the F-statistic from a standard ANOVA table. That is, φ^2 can be estimated as:

$$\widehat{\varphi}^2 = \frac{\sum_{i=1}^{k}(n_i/\bar{n})(\bar{X}_i - \bar{X}_{..})^2}{(N-k)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}.$$

Under the assumptions that the k groups of observations are normally distributed with common variance σ^2, $(\bar{n}/(k-1))\,\hat{\varphi}^2$ follows a non-central F distribution with $k-1, N-k$ degrees of freedom (where $N = \sum_{i=1}^{k} n_i$) and non-centrality parameter $\lambda^2 = \bar{n}\varepsilon^2$.

The critical region of the test is given by $\{\hat{\varphi}^2 < ((k-1)/\bar{n})F_{k-1,N-k;\,\alpha}(\bar{n}\varepsilon^2)\}$. Equivalence will be declared if the observed statistic $\hat{\varphi}^2$ falls in the critical region. Lack of equivalence will be declared if $\hat{\varphi}^2$ falls outside the critical region.

For this study, you need to use a mixed-effect model to obtain estimates of φ^2. With a pre-define equivalence margin of 0.5, the F test for equivalence can be conducted using the following SAS code:

```
proc mixed data=stab plots(only)=(studentpanel);
  class tp animal;
  model response = tp dose rep / ddfm=KR(firstorder);
  random int / sub=animal type=VC;
  ods output lsmeans=means diffs=diffs tests3=type3;
run;

data equiv;
  length trt $20.;
  set type3(where=(effect="TP"));
    ntp = 48;
    psi_sq=fvalue*(numdf/ntp);
    eps = round(quantile('NORMAL',.70), 0.1);
    c   =(numdf/ntp)*finv(.05, numdf, dendf, ntp*eps**2);
    eq  = (psi_sq < c1);

  keep eps numdf dendf fvalue psi_sq c eq;
run;
```

Where in the PROC MIXED step:

- ANIMAL is animal ID;
- TP represents time point (0, 3, 6, or 12 months),
- RESPONSE is the measured immune response;
- DOSE represents dose level (6 doses in total);
- REP indicates the replication.

The DATA step calculates the estimates of φ^2, where:

- NTP is the number of independent observations per time point;
- PSI_SQ is the estimate of φ^2;
- FVALUE is the SAS function returning a quantile of the F-distribution;
- EPS is the equivalence margin for the F-test corresponding to the original margin of 0.2 (see Wellek, page 164);
- C is the upper bound of the critical region, to be compared to the observed F-statistic;
- NUMDF is the F-statistic's numerator degrees of freedom (number of time points – 1);
- DENDF is the F-statistic's denominator degrees of freedom (estimated using the Kenward-Rogers method within "the MIXED procedure").

Table 1 summarizes the corresponding results of the F-test for equivalence for sample data.

NUMDF	DENDF	FVALUE	PSI_SQ	EPS	C	Assessment
3	186	1.15	0.07172	0.5	0.099044	Equivalent

Table 1. Results from the F-test for Equivalence

The conclusion of equivalence can be declared, because the estimate of φ^2 falls in the critical region (0.07172 < 0.099044).

CONCLUSION

This paper discussed equivalence tests for different study designs: one is for designs with paired observations, the other one is for designs with multiple treatment arms. Both study designs are widely used in different industries.

The Two One-Sided Test (TOST) is widely used to test pair-wise equivalence. The 90% CI of the difference is used rather than 95% CI because testing for equivalency is comprised of the two null hypotheses (TOST) stated above.

In clinical trials, according to FDA guidance, for the analysis of replicated studies, mixed-effect models should be used to obtain estimates of the mean differences and the 90% CIs for the differences.

The F-test is "global" in that it is not conducted pair-wise (such as TOST), but rather is a single, multiple degree-of-freedom, test that determines whether any group, within a set of k groups, is non-equivalent to any of the other group. The assumptions on the distributions of the observations must be satisfied in order to ensure that the F-test is a valid procedure.

In practice, the log-normal distribution is often assumed for data from clinic trials. You can perform equivalence analyses based on the log-transformed data and use the back-ward transformed 90% confidence interval to make conclusions.

REFERENCES

1. Wellk, S. 2003. *Testing Statistical Hypotheses of Equivalence*. New York: Chapman & Hall/CRC.
2. Ocana, J., Pilar O., Sanchez A. and Carrasco J. 2008. "On Equivalence and Bioequivalence Testing." *SORT*, 32 (2) July-December 2008, 151–176.
3. Hoenig, J.M. and Heisey, D.M. 2001. "The Abuse of Power: the Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician*, 55:19-24.
4. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Biologics Evaluation and Research, JANUARY 2001.

ACKNOWLEDGMENTS

We would like to thank Statistics Team in McDougall Scientific Ltd., Hong Chen, Jim Wang, and Hao Xu, for their support and advice.

RECOMMENDED READING

- *SAS/STAT® 9.2 User's Guide, The FREQ Procedure*

- *SAS/STAT® 9.2 User's Guide, The TOST Procedure*
- *SAS/STAT® 9.2 User's Guide, The MIXED Procedure*
- *SAS/STAT® 9.2 User's Guide, The POWER Procedure*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Fei Wang
McDougall Scientific Ltd.
fwang@mcdougallscientific.com