Superiority, Non-Inferiority & Equivalence

Larry Bartkus
October 2015
The Basics

There are three ways to look at a comparison between two data sets:

- **Superiority** – Is one sample statistically better than another?

- **Non-Inferiority** – Is one sample not unacceptably worse than another?

- **Equivalence** – Are both samples the same as each other within a statistically significant margin (not unacceptably different from each other)?
Hypotheses

- **Superiority:**
  - $H_0: \mu_T = \mu_C$
  - $H_1: \mu_T \neq \mu_C$

- **Equivalence**
  - $H_0: (\mu_T - \mu_C) \geq \Delta$ and $(\mu_T - \mu_C) \leq -\Delta$
  - $H_1: -\Delta < (\mu_T - \mu_C) < \Delta$

- **Non-inferiority**
  - $H_0: (\mu_T - \mu_C) \leq -\Delta$
  - $H_1: (\mu_T - \mu_C) > -\Delta$

- Note that all tests can simply be done as a hypothesis test. The difference is whether we are looking for a difference on the high side, low side or on both sides of the criteria, such as specifications.
Why Use Non-inferiority?

From Schumi and Wittes’ Article:

Non-inferiority trials are clearly appropriate for some diseases and some treatments. When developing a new treatment to prevent tuberculosis, investigators might be willing to sacrifice some small amount of benefit (as reflected in the margin) for a simpler dosing schedule, fewer side effects, or other advantages, but they would be delighted if the new treatment were better than current therapies (hence no restriction on the upper bound).
Why Use Non-inferiority?

From Ewald’s Article:

From a clinician’s perspective, if a new drug is not better we at least want to know it is not worse than the old drug. Statisticians use different methods if they are testing only one end of the equivalence boundary. In effect clinicians do not care how far the good end of the 95% confidence interval goes, just as long as the ‘bad’ end is within an acceptable limit. For this reason most trials will use a non-inferiority analysis.
Non-inferiority Outcomes

From Allen and Seaman’s Article – A good comparison example

**a. Superiority shown:** The confidence bound of \((X - Y)\) does not include the 0 difference line, and the confidence bound exceeds the upper margin of non-inferiority. The result shows the superiority of Y. This result is identical to that of a hypothesis comparing X and Y in which superiority of Y is shown.

**b. Non-inferiority shown:** Non-inferiority is shown because the lower bound of \((X - Y)\) crosses the 0 difference line; although the upper bound does exceed the upper non-inferiority boundary, but the lower bound does not cross the lower non-inferiority boundary.

**c. Non-inferiority shown:** Here non-inferiority is shown because the confidence bounds for \((X - Y)\) are totally within the non-inferiority boundaries.

**d. Non-inferiority shown:** Non-inferiority is shown because, although the confidence bounds for \((X - Y)\) cross the 0 line, the lower bound does not cross the lower non-inferiority boundary.

**e. Inequality shown:** The lower bound of the confidence interval of \((X - Y)\) crosses the lower boundary of the inferiority margin.
Group Discussion

You have a peel apart type of package:

• Under what circumstances would you want to perform a Non-inferiority test?

• Under what circumstances would you want to perform a Superiority Test?

• Under what circumstances would you want to perform an Equivalency Test?
Equivalence Range

• Delta (Δ) is the largest difference in the treatment means that is clinically acceptable, so that a difference bigger than this would matter in practice.

• Choose and justify Δ a priori.

• -Δ and +Δ may be chosen asymmetrically with respect to zero.
Alpha Risk is the probability of rejecting a good lot. It can be expressed as:

\[ \text{Alpha} = 1 - \text{Prob}_{\text{acc}} \text{ (at the AQL level)} \]

- Alpha Risk (Type I Error)
  - Probability of rejecting a good lot
  - Probability of lots that meet AQL will not get accepted
  - Producer’s Risk
  - Consequences: Waste of labor, materials and decreased productivity due to rework, re-inspection, etc.
**Beta Risk** is the probability of accepting a bad lot. It can be expressed as:

\[ \text{Beta} = \text{Prob}_{\text{acc}} \text{ (at the LTPD level)} \]

- Beta Risk (Type II Error)
  - Probability of accepting a bad lot
  - Probability of lots that exceed LTPD will be accepted
  - Consumer’s Risk
  - It can also mean accepting a hypothesis that is false.
The Risks

- **Type I error**: Deciding the groups are different when they aren’t (the difference is due to random variation).

- **Type II error**: Not detecting a difference when there really is one.

- **P-value**: the probability of making a Type I, or alpha, error (deciding the groups are different when they really aren’t).
  - Choose the level of Type I error that can be tolerated; by convention, it is usually set at .05 (= 5% chance).
There are four possible outcomes to any decision made based on a hypothesis test: If the groups are different, and if the hypothesis is right or wrong.

- **Accept $H_0$: Groups are Same**
  - No Error
  - Type II Error

- **Reject $H_0$: Groups are Different**
  - Type I Error
  - No Error
Power and Sample Size

The power of a statistical test is the probability, given that $H_0$ is false, of obtaining sample results that will lead to its rejection.

“Power of a test” is clearly an important characteristic. To put it in other words, a powerful test is one that has a high probability of claiming that a difference exists when it really does.

-The larger the effect size, $d$, the greater the power.

-For any given effect size (other than zero) the larger the sample size, the greater the power of the test.
Power and Sample Size
Power and Sample Size

Hypothesis about: Test mean - reference mean (Difference)

What do you want to determine? (Alternative hypothesis)
Lower limit < test mean - reference mean < upper limit

Lower limit: -3
Upper limit: 3

Specify values for any two of the following:
Sample sizes:
Differences (within the limits): 1
Power values: 0.8
Standard deviation: 1.9

Options... Graph... Help OK Cancel
2-Sample Equivalence Test

Power for difference: Test mean - reference mean
Null hypothesis: Difference ≤ -3 or Difference ≥ 3
Alternative hypothesis: -3 < Difference < 3
α level: 0.05
Assumed standard deviation: 1.9

<table>
<thead>
<tr>
<th>Sample</th>
<th>Target</th>
<th>Difference</th>
<th>Size</th>
<th>Power</th>
<th>Actual Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>12</td>
<td>0.8</td>
<td>0.802754</td>
</tr>
</tbody>
</table>

The sample size is for each group.

Power Curve for 2-Sample Equivalence Test
The Power Curve

Power Curve for 2-Sample Equivalence Test

Assumptions
- Sample Size: 12
- Assumptions:
  - $\alpha$: 0.05
  - StDev: 1.9
Typically a 2-sample t-test is misused to determine equivalence. If the p-value is less than alpha, you conclude that the means significantly differ. But if the p-value is not less than alpha, you haven’t proven that the means are equal. You don’t have enough evidence to prove that they’re not equal.

The better and certainly more correct way to determine equivalence is to use the Two One-Sided Tests (TOST). This will evaluate the 2 data sets from both sides, the top and bottom to establish equivalence based upon some acceptable criteria.
Equivalence in Minitab

2-Sample Equivalence Test
Determine whether the means of 2 independent samples are equivalent. Use to compare a test product or process with a reference product or process.
Equivalence in Minitab
Equivalence in Minitab

Boxplot of Infusion Amount vs Device

Data

362 363 364 365 366 367 368 369 370
Equivalence in Minitab

Two-Sample Equivalence Test: Infusion Amount, Device

Method
Test mean = mean of A
Reference mean = mean of B
Equal Variances were not assumed for the analysis.

Descriptive Statistics

<table>
<thead>
<tr>
<th>Device</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12</td>
<td>366.64</td>
<td>2.0601</td>
<td>0.59471</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>367.16</td>
<td>1.6817</td>
<td>0.48896</td>
</tr>
</tbody>
</table>

Difference: Mean(A) - Mean(B)

<table>
<thead>
<tr>
<th>Difference</th>
<th>SE</th>
<th>95% CI</th>
<th>Equivalence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.52233</td>
<td>0.76770</td>
<td>(-1.8433, 0.78867)</td>
<td>(-3, 3)</td>
</tr>
</tbody>
</table>

CI is within the equivalence interval. Can claim equivalence.

Test
Null hypothesis: Difference ≤ -3 or Difference ≥ 3
Alternative hypothesis: -3 < Difference < 3
α level: 0.05

Null Hypothesis DF T-Value P-Value
Difference ≤ -3 21 3.2274 0.002
Difference ≥ 3 21 -4.5882 0.000

The greater of the two P-Values is 0.002. Can claim equivalence.

Equivalence Test: Mean(A) - Mean(B)
Equivalence in Minitab

- Another Example of Equivalence from Patrick Runkel’s Article

**Equivalence Test: Mean(New) - Mean(Current)**

Test
Null hypothesis: Difference ≤ -0.4 or Difference ≥ 0.4
Alternative hypothesis: -0.4 < Difference < 0.4
α level: 0.05

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>DF</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference ≤ -0.4</td>
<td>12</td>
<td>1.3717</td>
<td>0.098</td>
</tr>
<tr>
<td>Difference ≥ 0.4</td>
<td>12</td>
<td>-2.5646</td>
<td>0.012</td>
</tr>
</tbody>
</table>

The greater of the two P-Values is 0.098. Cannot claim equivalence.
Summary

• To choose between an equivalence test and a standard t-test, consider what you hope to prove or demonstrate. Whatever you hope to prove true should be set up as the alternative hypothesis for the test and require the burden of proof. Whatever you deem to be the less harmful incorrect assumption to make should be the null hypothesis. If you’re trying to rigorously prove that two means are equal, or that a mean equals a target value, you may want to use an equivalence test rather than a standard t-test. (From Patrick Runkel’s Article)
References

• Superiority, Equivalence and Non-Inferiority by I. Elaine Allen and Christopher A. Seaman (Quality Progress Feb. 2007)
• Comparing Two Treatments Using Equivalence Tests or Non-inferiority Tests by David Jackson and Lily Jeng (April 2007)
• Minitab 17 Training Material – Statistical Tools for Medical Devices: Additional Topics (2014 Minitab Inc.)
• Making sense of equivalence and non-inferiority trials by Ben Ewald, University of Newcastle, New South Wales
• Through the looking glass: understanding non-inferiority by Jennifer Schumi and Janet T. Wittes
Any Questions?