Approach

- We’ll start with a review of significance tests and then transition to equivalence tests.
- We’ll use the two-sample $t$ test for means for this discussion.
- Understand that these methods also apply to:
  - other parameters (e.g. standard deviations, proportions, counts, etc.)
  - problems involving one, two, and many populations.
Review: Two-Sample $t$ Test

- The hypotheses for the two-sample location problem are

$$ H_0 : \mu_1 = \mu_2 $$
$$ H_A : \mu_1 \neq \mu_2 $$

or in terms of $\Delta \mu = \mu_1 - \mu_2$

$$ H_0 : \Delta \mu = 0 $$
$$ H_A : \Delta \mu \neq 0 $$

- This test is a significance test because its purpose is to demonstrate $H_A$ - that there is a significant difference between the two population means.
- So acceptable decisions are limited to either rejecting $H_0$ in favor of $H_A$ or reserving judgement, i.e. we never accept $H_0$.

Two-Sample $t$ Test Example

Example: An experiment was performed to test for a difference between two population means. The two samples yielded the following results:

$$ n_1 = 8, \bar{x}_1 = 18.8, s_1 = 1.5 $$
$$ n_2 = 10, \bar{x}_2 = 15.6, s_2 = 2.4 $$

Perform the hypothesis test using $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$ and construct the 95% confidence interval for the bias between the population means.
Two-Sample \( t \) Test Example

Solution: Using MINITAB Stat> Basic Statistics> 2-Sample \( t \) the test statistic is \( t = 3.46 \) with \( p = 0.004 \) and the confidence interval is \( P(1.23 < \Delta \mu < 5.17) = 0.95 \). There are three ways to decide whether to reject \( H_0 \) or not.

- \( (t = 3.46) > (t_{0.025} = 2.12) \) so reject \( H_0 \)
- \( (p = 0.004) < (\alpha = 0.05) \) so reject \( H_0 \)
- The confidence interval does not contain zero so reject \( H_0 \)

Design of Hypotheses

- To test the hypotheses

\[
H_0: \text{Something ordinary happens}
\]

versus

\[
H_A: \text{Something extraordinary happens,}
\]

"the extraordinary claim requires extraordinary evidence." - Carl Sagan

- The extraordinary claim is always \( H_A \).
- \( H_A \) comes first. \( H_0 \) is just its complement.
- There is no special interest in or value to \( H_0 \).
- Sagan says:
  - The extraordinary claim \( H_A \) is what we want to demonstrate.
  - The data must be very strong to reject \( H_0 \) and accept \( H_A \).
  - There is no opportunity to accept the ordinary claim \( H_0 \).
  - Put the burden of proof on the data. That is, if the data are weak or insufficient then we can’t reject \( H_0 \).
Design of Hypotheses

Example: Formulate hypotheses to test the claim that there are (outer space) aliens here among us. What evidence is required?

Solution: The extraordinary claim is that there are aliens here among us so

\[ H_0: \text{there are no aliens here among us} \]
\[ H_A: \text{there are aliens here among us} \]

To reject \( H_0 \) in favor of \( H_A \):

- Extraordinary evidence - we must meet a walking talking alien, be allowed to perform some physiology/medical tests on him/her/it, see their space ship, visit their planet to confirm that there are others like him/her/it, etc.
- Weak evidence - the evidence that the aliens are here is that we can’t see them because they’re too smart to get caught.

Design of Hypotheses

Exercise: Formulate the hypotheses to test the claim "there are no mice in my house." What data would be required to support the claim?

Solution:
Run the Experiment
Collect sample data and calculate a test statistic that approximates the hypothesized parameter value. For our two-sample *t* test example we’re using $\Delta \bar{x}$ to estimate $\Delta \mu$.

Hypothesis Test Interpretation
The hypothesis test can be interpreted:
- by comparing the *t* test statistic to $t_{a/2}$
- by comparing the *p* value to $\alpha$
- by checking if the confidence interval for $\Delta \mu$ contains zero or not.

So reject $H_0 : \Delta \mu = 0$ when
- $|t| > t_{a/2}$ or
- $p < \alpha$ or
- when the confidence interval for $\Delta \mu$ does not contain zero.
Equivalence Test

- The goal of a significance test is to show that there is a statistically significant difference between the population means.
- The goal of an equivalence test is to show that the population means are equivalent to each other.
- The hypotheses for the two-sample equivalence test are:
  \[ H_0 : \mu_1 \neq \mu_2 \]
  \[ H_A : \mu_1 = \mu_2 \]
  or in terms of \( \Delta \mu = \mu_1 - \mu_2 \)
  \[ H_0 : \Delta \mu \neq 0 \]
  \[ H_A : \Delta \mu = 0 \]

- Notice that these hypotheses are just the inverted hypotheses from the significance test.
- How do we perform the equivalence test?

Why Not Just Accept \( H_0 \)?

If we retain the original significance test hypotheses \( H_0 : \Delta \mu = 0 \) versus \( H_A : \Delta \mu \neq 0 \) what’s wrong with accepting \( H_0 \)?

- Using the confidence interval interpretation we would accept \( H_0 \) if the confidence interval contains zero.
- The confidence interval has the form
  \[ P(\Delta \bar{x} - \delta < \Delta \mu < \Delta \bar{x} + \delta) = 1 - \alpha \]
  where the confidence interval half-width is
  \[ \delta = t_{\alpha/2} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]

- If we run a bad experiment so that \( \sigma \) is very large or if the sample sizes are too small then the \( \delta \) will be very large and the confidence interval will very wide and almost certainly contain zero.
- Absence of evidence is not evidence of absence.
- This approach - accepting \( H_0 : \Delta \mu = 0 \) - does not put a sufficient burden of proof on the data.
- The solution to the problem is an equivalence test.
Equivalence Test Procedure

- When the purpose of a test is to demonstrate the equality of two treatment means we perform an equivalence test using hypotheses:
  \[ H_0 : \Delta \mu \neq 0 \]
  \[ H_A : \Delta \mu = 0 \]

Note that what we want to demonstrate - the extraordinary claim - is \( H_A \).

- It is impossible to show that \( H_A \) is exactly true because there could be a practically small and insignificant bias between \( \mu_1 \) and \( \mu_2 \) so we use different forms for these hypotheses - forms that formally consider the possibility of such a bias:
  \[ H_0 : |\Delta \mu| \geq \delta \]
  \[ H_A : |\Delta \mu| < \delta \]

where \( \delta \) is called the limit of practical equivalence (LOPE).

Equivalence Test Procedure

- LOPE is the largest value of \( \Delta \mu \) for which you would consider the two population means to be practically equivalent to each other.
- LOPE must be sufficiently small so that the claim that \( \Delta \mu = 0 \) is, for all practical purposes, still justified.
- For this reason the value of the LOPE must be chosen by or be satisfactory to the process owner.
- When we perform the equivalence test, if the data support \( H_A : |\Delta \mu| < \delta \) then we say that \( \mu_1 \) and \( \mu_2 \) are practically equivalent to each other.
The equivalence test is performed using two one-sided tests of means (TOST). The absolute values in the original equivalence test hypotheses are broken up into two separate tests:

- $H_{01} : \Delta \mu \leq -\delta$ versus $H_{A1} : \Delta \mu > -\delta$
- $H_{02} : \Delta \mu \geq \delta$ versus $H_{A2} : \Delta \mu < \delta$

To perform the equivalence test by the confidence interval method construct the $(1 - 2\alpha)100\%$ confidence interval for $\Delta \mu$. If the confidence interval falls completely inside of the interval $\pm \delta$ then reject $H_0$ in favor of $H_A$.

To be conservative some experts advise using the $(1 - \alpha)100\%$ confidence interval instead.

Note that the confidence interval can be interpreted in two ways:
- in terms of the significance test by checking the location of the confidence interval with respect to 0
- in terms of the equivalence test by checking the location of the confidence interval with respect to the $\pm \text{LOPE}$ values.
**TOST Method**

**Example:** Suppose that the goal of our example problem was to demonstrate that $\mu_1$ and $\mu_2$ were practically equivalent and that before the data were collected the process owner chose LOPE to be $\delta = 3$. Interpret the confidence interval in terms of the significance test and the equivalence test.

**Solution:** The plot below shows the confidence interval for $\Delta \mu$.

- **(Significance test)** The confidence interval does not include 0 so we must reject $H_0 : \Delta \mu = 0$ in favor of $H_A : \Delta \mu \neq 0$, i.e. there is a statistically significant difference between $\mu_1$ and $\mu_2$.

- **(Equivalence test)** The confidence interval does not fall completely within the interval $\delta = \pm 3$ so we cannot reject $H_0 : |\Delta \mu| \geq \delta$, i.e. we do not have evidence for equivalence.

![Graph showing the confidence interval for $\Delta \mu$.](image)

---

**TOST Method**

**Exercise:** Use the TOST confidence interval method for equivalence tests to interpret cases A to D in the plot below. Also interpret the intervals in terms of significance tests.

![Graph with intervals labeled A, B, C, D.](image)
Sample Size Calculation

- The sample size required for the two-sample t test to reject $H_0 : \mu_1 = \mu_2$ with power $\pi = 1 - \beta$ when the difference between the means is $\Delta \mu$ is given by
  \[
  n = 2 \left( t_{\alpha/2} + t_{\beta} \frac{\sigma_\varepsilon}{\Delta \mu} \right)^2
  \]

- The sample size required for the two-sample equivalence test with limit of practical equivalence $\delta$ and true bias between the means $\Delta \mu$ is
  \[
  n = 2 \left( t_{\alpha} + t_{\beta} \frac{\sigma_\varepsilon}{\delta - \Delta \mu} \right)^2
  \]

Example: Determine the sample size required for the two-sample equivalence test if the LOPE is $\delta = 2$ and $\Delta \mu = 0.2$ with 90% power and $\sigma_1 = \sigma_2 = 2$.

Solution: With $\delta = 2$ the hypotheses to be tested are

$H_{01} : \Delta \mu \leq -2$ versus $H_{A1} : \Delta \mu > -2$

$H_{02} : \Delta \mu \geq 2$ versus $H_{A2} : \Delta \mu < 2$.

To obtain 90% power we have $\beta = 1 - 0.90 = 0.10$ so the sample size is

\[
\begin{align*}
n &= 2 \left( t_{\alpha} + t_{\beta} \frac{\sigma_\varepsilon}{\delta - \Delta \mu} \right)^2 \\
&\approx 2 \left( \frac{1.645 + 1.282}{2 - 0.2} \right)^2 \\
&\approx 22
\end{align*}
\]

Further iterations of the t values give $n = 24$. 

Mathews Malnar & Bailey, Inc., Equivalence Tests
Solution using MINITAB (V17)
Stat> Power and Sample Size> Equivalence Tests> 2-Sample:

Superiority Test

- The goal of a superiority test is to demonstrate that one treatment mean is greater than the other. Superiority tests are one-sided so a sense of direction, either larger is better or smaller is better, must be specified.
- If a larger response is better, then to demonstrate that \( \mu_1 \) is superior to \( \mu_2 \) the hypotheses are

\[
H_0 : \mu_1 = \mu_2 \text{ or } \Delta \mu = 0 \\
H_A : \mu_1 > \mu_2 \text{ or } \Delta \mu > 0
\]

These hypotheses are identical to those of the one-sided significance test.
- If a larger response is better, then reject \( H_0 \) if the confidence interval for \( \Delta \mu \) is completely to the right of and does not contain zero.
Noninferiority Test

- The goal of a noninferiority test is to demonstrate that one population’s mean is not inferior to another population’s mean or, if it is inferior, it is not inferior by very much. Noninferiority tests are one-sided so a sense of direction, either larger is better or smaller is better, must be specified.
- To perform the noninferiority test a limit or margin of noninferiority \( \delta \) must be specified. This is the largest value that \( \mu_1 \) could be worse than \( \mu_2 \) by and still be considered noninferior to \( \mu_2 \). If a larger response is better, then the value of the limit must be negative.
- If a larger response is better, then to demonstrate that \( \mu_1 \) is noninferior to \( \mu_2 \) the hypotheses are

\[
H_0 : \Delta \mu = -\delta \\
H_A : \Delta \mu > -\delta
\]

where \( \Delta \mu = \mu_1 - \mu_2 \). Reject \( H_0 \) if the confidence interval for \( \Delta \mu \) falls completely above the noninferiority limit at \( -\delta \).

Exercise: Interpret the confidence intervals in the context of noninferiority tests where larger is better.

Warning: Situation d presents a minor contradiction. Technically it shows noninferiority (the confidence interval falls completely to the right of \( -\delta \)); however, it also shows that \( \mu_1 \) is statistically significantly smaller than \( \mu_2 \).

Exercise: Sometimes a noninferiority claim can be upgraded to a superiority claim. For which of the cases above can this be done and why?
References