Two-sample t-test for means

Use case

We have two independent (i.e., disjoint) populations. We want to determine whether there is a statistically significant difference between the means of each population. We don't know with certainty the standard deviation of one or both populations,
σ_1 and
σ_2, so we estimate them using the sample standard deviations,
s_1 and
s_2. We have a sample of size
n_1 from population 1, and a sample of size
n_2 from popluation 2. Population 1 yields sample mean
$\bar{x}_1$, and population 2 yields sample mean
$\bar{x}_2$.

Preconditions

To apply a two-sample t-test for means, one should verify the following conditions:

1. If
   $n_1 + n_2 < 15$, it is critical that the populations are normally distributed.
2. If
   $n_1 + n_2 \geq 15$, one may proceed in the absence of outliers or strong skewness.
3. If
   $n_1 + n_2 \geq 40$, procedures are generally robust.

One can enhance robustness by planning
$n_1 \approx n_2$.

Hypotheses

Null hypothesis,
H_0**:**
$\mu_1 - \mu_2 = 0$ (i.e., there is no difference between the two populations.)

Alternative hypothesis,
H_a**:**
$\mu_1 - \mu_2 \neq / > / < 0$ (i.e., there is a difference between the two populations.)

Test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

p-value

Where T is distributed according to the
$t(\min(n_1 - 1, n_2 - 1))$ distribution *(conservative approach)*,

1. For
   $H_a : \mu_1 - \mu_2 < 0$:
   $p = P(T \leq t)$

2. For
   $H_a: \mu_1 - \mu_2 > 0$:
   $p = P(T \geq t)$
3. For
   $H_a: \mu_1 - \mu_2 \neq 0$:
   $p = 2P(T \leq -|t|)$

Conservative approach

By choosing the smaller of
n_1 and
n_2 as the basis for our degrees of freedom, we take a *conservative approach*. As a result, our p-values are slightly larger than the true p-value.

p-value (Python)

1. For
   $H_a: \mu_1 - \mu_2 < 0$: `p = stats.t.cdf(t, df=min(n1-1,n2-1))`
2. For
   $H_a: \mu_1 - \mu_2 > 0$: `p = 1 - stats.t.cdf(t, df=min(n1-1,n2-1))`
3. For
   $H_a: \mu_1 - \mu_2 \neq 0$: `p = 2 * stats.t.cdf(-abs(t), df=min(n1-1,n2-1))`

Associated confidence interval

$$\text{C\% confidence interval} = (\bar{x}_1 - \bar{x}_2) \pm t^\star \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
$$\text{choose } t^\star \text{ s.t. area on } t(\min(n_1 - 1, n_2 - 1)) \text{ distribution from } (-t^\star, t^\star) = \text{C}$$

Because we employ a *conservative approach*, this confidence interval will be slightly wider than the true confidence interval.

Related tests

If you want to perform inference on the mean of a single population, a one-sample t-test for means is more appropriate.

If you have access to both of the population standard deviation values, a two-sample z-test for means is more appropriate.

Considerations

A two-sample t-test conducted at confidence level
α will reject the null hypothesis if and only if the value corresponding to the null hypothesis,
0, is completely outside of the
$C = 1 - \alpha$ t-confidence interval for the true difference between means.