Independent Samples t Test

What it is

The independent samples t focuses on the difference between two population means. The samples the data come from must be independent, meaning observations in one sample are unrelated to those made in the second (compare to the *Paired Samples t Test*). With this test we usually set the null hypothesis to equality between the two population means[1]. The test then measures the empirical evidence against equality (how much evidence we have that the population means differ).

[1] $H_0 : \mu_A = \mu_B$ or $H_0 : \mu_A - \mu_B = 0$

When to use it

The independent samples t test is useful when data come from two simple random samples of distinct populations and the samples were drawn independently from each other. For the t approximation to work well, each sample needs to meet the requirements of the CLT[2].

[2] Each sample should be approximately normally distributed or large (rule of thumb: $n \geq 30$ for each).

How to use it

In R, the independent samples t test is performed using the `t.test()` function. There are two ways to use this method: entering data from the two samples separately, or using what R calls a "formula" to separate the observed values by a second variable that indicates which group each observation comes from. The second is used when both the data and grouping variable are in the same data set. It is probably the most common way to perform a t test, so this is what we discuss below.

To compare average BMI between Missouri men and women, we enter the formula `bmi ~ sex` into the `t.test()` function[3]:

[3] This formula can be read as: "BMI as a function of sex" or "separate BMI by values of sex variable."

```
t.test(bmi ~ sex, data=mydat, alternative='two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by sex
## t = 2.4232, df = 1844.7, p-value = 0.01548
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1384212 1.3137348
## sample estimates:
##   mean in group Male mean in group Female
##             28.95919             28.23311
```

The output contains the t statistic, degrees of freedom, and p-value for the hypothesis test, as well as the 95% confidence interval for the difference between population means and the sample means within each of the groups.

In this example, we see males average a 28.95 BMI while females average 28.23. This difference is small but we have a pretty large sample, so the t statistic is about 2.4 and the p-value is 0.015 - there's only about a 1 or 2 in 100 chance that such a large difference would occur by chance in a sample of this size if there was no difference in population averages. Additionally, the confidence interval tells us we are pretty sure the average male BMI is about 0.1 to 1.3 points higher than the average female BMI.

A few important notes: First, R orders the groups numerically or alphabetically. In our case, R is comparing BMI between the sexes as $\mu_{Male} - \mu_{Female}$ since the 1 coded as "Male" comes first numerically[4]. This is important because the t statistic also follows this ordering, so the positive t statistic tells us that in our sample males have a higher average BMI than females ($\bar{Y}_{Male} - \bar{Y}_{Female} > 0$). Similarly, the positive confidence interval indicates the male average BMI is greater than the female average. To be sure about the order of the groups, they match the order of the sample means at the end of the `t.test()` output. The second mean is being subtracted from the first mean.

[4] This example is a little tricky since it is not alphabetical. However, the `sex` variable is technically a factor with values 1 and 2 labelled as "Male" and "Female" respectively. You can see this by running `str(mydat$sex)`.

Second, we can change `alternative` to `'less'` or `'greater'` for one-sided tests. Again, knowing the order R will place the groups is especially important in this case. In our example, setting `alternative='less'` will test whether males have an average BMI *lower* than females, while `alternative='greater'` will test whether males have an average BMI *higher* than females.

Why it works

When comparing two means, the difference between them is a straightforward choice and easiest mathematically (compared to, say, ratios). For two independent sample means, the standard error of the difference is $SE(\bar{Y}_A - \bar{Y}_B) = \sqrt{SE(\bar{Y}_A)^2 + SE(\bar{Y}_B)^2}$.

Under the conditions of the CLT and assuming $H_0 : \mu_A - \mu_B = 0$ is true,

$$t^* = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_A - \mu_B)}{SE(\bar{Y}_A - \bar{Y}_B)} = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

follows a t distribution. The degrees of freedom here follow a complicated formula[5], but they can be estimated reasonably well by the much simpler formula $df = \min\{n_A, n_B\} - 1$, the degrees of freedom in the smaller of the two samples.

[5] The best approximation is

$$df = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{(s_A^2/n_A)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}}$$