Independent samples t test

2020-03-29

What it is

The independent samples t test the difference between two population means. The samples the data come from must be independent, meaning when 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]. Our test then measures the amount of evidence we have 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. Each sample needs to meet the requirements of the CLT: each 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.

Suppose we want to compare average BMI between men and women. The R formula for this is `bmi ~ sex`. We enter this into the `t.test()` function:

```
t.test(bmi ~ sex, data = dat, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by sex
## t = -1.2521, df = 498, p-value = 0.2111
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -1.6015940  0.3547947
## sample estimates:
## mean in group female   mean in group male
##             26.61093             27.23433
```

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.

A few important notes: First, R orders the groups numerically or alphabetically, so in the case above R is comparing BMI between the sexes as $\mu_{female} - \mu_{male}$ since "female" comes first alphabetically.[2] This is important because the t statistic also follows this ordering, so the negative t statistic tells us that in our sample females have a lower average BMI than males ($\bar{Y}_{female} - \bar{Y}_{male} < 0$). 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.

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 the example above, setting `alternative='less'` will test whether females have an average BMI *lower* than males, again due to alphabetical ordering of the groups.

[2] This assumes the group variable is a character string. If the group variable is a factor, R will use the internally coded order of the factor levels. R also defaults to an alphabetically ordering when creating factors, but this can be changed by user input. It is important to check both the class and order (if a factor) of your group variable.

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[3], 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.

[3] The best approximation is

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