Independent samples t test

Date: 2017-11-01

What it is

The independent samples t test compared the means in two samples when observations made in one sample are unrelated to those made in the second (compare to the *Paired samples t test*). We usually use this test to test for the equality of the two means in the populations from which the samples were drawn.[1]

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. Similar to the CLT requirements for the confidence interval for μ, each sample should be large (rule of thumb: $n \geq 30$ for each).

How to use it

In R, the independent samples t test is performed using the function `t.test()`. 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 values by a second variable that indicates which sample each observed value comes from. The second is used when both the data and grouping variable are in the same data set. This is also 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 = d, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by sex
## t = -1.0448, df = 472.27, p-value =
## 0.2967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7810561  0.5445458
```

[1] $H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 \neq \mu_2$

```
## sample estimates:
## mean in group female   mean in group male
##             27.43951             28.05777
```

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 numericall 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 we can see that females have a lower BMI because the t statistic is negative. 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, seeting `alternative='less'` will test whether females have an average BMI *lower* than males, again due to alphabetical ordering of the groups.

Why it works

When comparing two means, the difference between them is an obvious choice. Probability theory tells us for independent random variables X and Y, $Var(X - Y) = Var(X) + Var(Y)$.

Using X as $\bar{Y}_1$ and Y as $\bar{Y}_2$, $Var(\bar{Y}_1 - \bar{Y}_2) = Var(\bar{Y}_1) + Var(\bar{Y}_2)$. Since both are sample means, $Var(\bar{Y}_1 - \bar{Y}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. Taking the square root gives the standard error.

Now, assuming the CLT conditions hold, $\bar{Y}_1 - \bar{Y}_2/SE(\bar{Y}_1 - \bar{Y}_2)$ 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 $df = \min\{n_1, n_2\} - 1$, the degrees of freedom in the smaller of the two samples.

[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, so it is important to check both the class and order (if a factor) of your group variable.

[3] $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{s_2^2/n_2)^2}{n_2 - 1}}$