[STAT-08] The t tests

Miguel-Angel Canela
Associate Professor, IESE Business School

An example

The logic of hypothesis testing is not obvious for the beginner. But practice teaches us how efficient is to use the same argument in many different situations. Building upon that experience, I skip the theoretical discussion and, instead of a formal definition, I start with an example.

Example 1. The `lwages` data set includes data on wages in years 1980 and 1987. The sample size is $n = 545$. The variables are: (a) `nr`, an identifier, (b) `lwage0`, wages in 1980, in thousands of US dollars and (c) `lwage7`, the same for 1987. The wages come in log scale. Do these data support that there has been a change in the wages? Note that we don't care about individual changes, but about the average change.

To explore this point, I introduce a new variable X, corresponding to the difference between these two years (1987 minus 1980). Now, to answer the ablove question, I search for evidence that the mean of X is different of zero. My first approach is based on a confidence interval. The basic information is

$$\bar{x} = 0.473, \qquad s = 0.606, \qquad n = 545.$$

With the appropriate t factor ($t_{0.025}(544) = 1.964$), I get the 95% confidence limits 0.422 and 0.524. Since this interval does not contain zero, it can be concluded (95% confidence) that there has been a change. It is then said that the mean difference $\bar{x} = 0.473$ is significant or, more specifically, that it is significantly different from zero.

¶ Source: F Vella & M Verbeek (1998), Whose wages do Unions raise? A dynamic model of unionism and wage rate determination for young men, *Journal of Applied Econometrics* **13**, 163–183.

The one-sample t test

Let me now tell you the story in a different way. I denote by μ the population mean of the wage increase (in log scale). Then, the conclusion of the above argument can be stated saying that I tested the **null hypothesis** $H_0 : \mu = 0$. Applied to the actual data, the test rejected H_0. So, I concluded that $\mu \neq 0$, with 95% confidence.

An alternative way to perform the test is based on the **test statistic**

$$T = \frac{\bar{X}}{S/\sqrt{n}}\,,$$

which, under the null (i.e. assuming that H_0 is valid) and an implicit normality assumption, follows a $t(n-1)$ distribution. So, we call this a t **test**. The absolute value of this statistic is compared with the **critical value** $t_{0.025}$, which corresponds to a 95% interval. If the critical value is exceeded, H_0 is rejected, with 95% confidence. We say then that the t value is significant.

Although the 95% level is a standard, we can change it by replacing $t_{0.025}$ by the critical value t_α, corresponding to the confidence level $1 - 2\alpha$. Mind that the use of other levels than the usual 95%

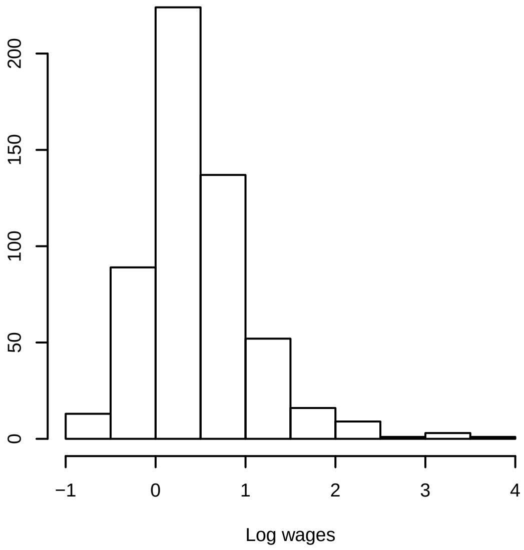

Figure 1. Distribution of log wage differences (Example 1)

has to be justified. Here, the value

$$t = \frac{0.473}{0.606/\sqrt{545}} = 18.22$$

exceeds the critical value 1.964. Again, we reject H_0 and conclude that there is a change in mean wages.

By replacing $\bar{x}$ by $\bar{x} - \mu_0$, this test can be applied to a null $H_0 : \mu = \mu_0$, in which μ_0 is a prespecified value. Without normality, the same methods are approximately valid for big samples. It is generally agreed that $n \geq 50$ is big enough for that. Of course, for big samples, the difference between the critical value $t_{0.025}$ and $z_{0.025} = 1.960$ becomes irrelevant.

The result of this test is usually presented in terms of a ***p*-value**. This is the 2-tail probability P associated, under the null, to the actual value of the t statistic. It is taken as a measure of the extent to which the actual results are significant (the lower the p-value, the higher the significance). With a 95% confidence level, we consider that there is significance when $P < 0.05$. In the example,

$$P = \mathrm{p}\big[|T| > 18.22\big] = 1.54 \times 10^{-58}.$$

The two-sample t tests

I consider now the null $H_0 : \mu_1 = \mu_2$, where μ_1 and μ_2 are the means of two distributions. The **two-sample *t* test** applies to two independent samples of these distributions. Both distributions are assumed to be normal, but this can be relaxed for big samples, as in the one-sample test.

In the simplest variant, it is assumed that the variance is the same for the two distributions compared ($\sigma_1 = \sigma_2$). If this assumption is not valid, we use a second variant, a bit more involved. Because of this extra complexity, textbooks frequently present a complete justification of the first variant, giving less detail about the second variant. Nevertheless, this complexity is irrelevant with a computer at hand, since you can run both variants in seconds. In practice, the p-values of the two tests are very close, except (possibly) when $\sigma_1 << \sigma_2$ and the sample sizes are very different.

The t test for equal variances

Assuming equal variances, we have two independent samples, of sizes n_1 and n_2, from the distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$. The mean difference $\bar{X}_1 - \bar{X}_2$ is normally distributed (I do not prove this) and, due to the properties of the sample mean, satisfies

$$\mathrm{E}\big[\bar{X}_1 - \bar{X}_2\big] = \mu_1 - \mu_2, \qquad \mathrm{var}\big[\bar{X}_1 - \bar{X}_2\big] = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

So,

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\,\sqrt{(1/n_1) + (1/n_2)}}$$

has a standard normal distribution. σ is unknown in real-world data analysis, but, under the equal variances assumption, we have here two unbiased estimators S_1^2 and S_2^2. The weighted average

$$S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}\,S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\,S_2^2,$$

called the **pooled variance** is also an unbiased estimator of σ^2. It can be proved that, under the null, the t statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\,\sqrt{(1/n_1) + (1/n_2)}}\,,$$

has a $t(n_1 + n_2 - 2)$ distribution. So, we take it as the test statistic here. The p-value is 2-tail area associated to the actual value of the t statistic in this distribution.

Alternative version

If it is not assumed that $\sigma_1 = \sigma_2$ (nor that they are different), we can use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}\,, \qquad \mathrm{df} = \frac{\big(S_1^2/n_1 + S_2^2/n_2\big)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}\,.$$

Under the null, the distribution of T can be approximated by a Student t, with the degrees of freedom given by the second formula. This is called the **Satterthwaite approximation**. The number of degrees of freedom is rounded when the test is done manually.

Example 2. In a cross-cultural study, the influence of gender, marital status and country citizenship on different aspects of well-being has been examined, testing the uniformity within the "Latin" world and comparing the variance due to the country effect with those due to the gender and marital status effects. The data were collected on a sample of managers following part-time MBA programs in nine Latin countries. The `jobsat1` data set contains data on job satisfaction (average of a 12-item Likert scale) for a subsample covering three countries, Chile (CH), Mexico (ME) and Spain (SP).

The sample size is $n = 423$, and the group sizes $n_0 = 121$, $n_1 = 111$ and $n_2 = 191$, for Chile, Mexico and Spain, respectively. The group statistics are reported in Table 1.

TABLE 1. Group statistics (Example 2)

Statistic	Chile	Mexico	Spain	Total
Size	121	111	191	423
Mean	4.158	4.413	4.162	4.227
Stdev	0.902	0.865	0.814	0.858

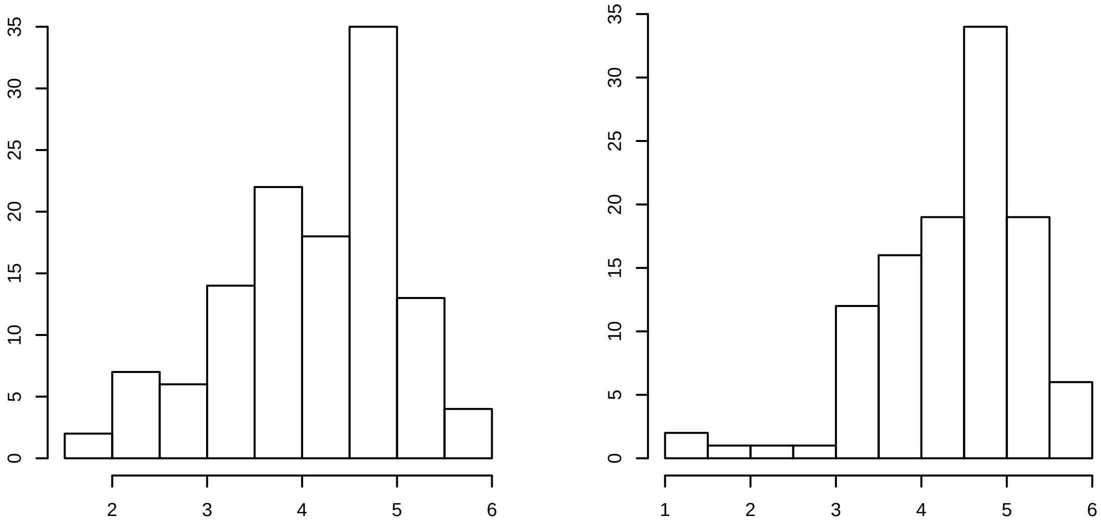

Figure 2. Distribution of job satisfaction in Chile and Mexico (Example 2)

I apply the two-sample t test to the mean difference between Chile and Mexico. We have

$$\bar{x}_1 = 4.158, \qquad s_1 = 0.902, \qquad n_1 = 121, \qquad \bar{x}_2 = 4.413, \qquad s_2 = 0.865, \qquad n_2 = 111.$$

Then, in the equal-variances version,

$$s = \sqrt{\frac{120 \times 0.902^2 + 110 \times 0.865^2}{230}} = 0.884, \qquad t = \frac{4.158 - 4.413}{0.884\sqrt{1/121 + 1/111}} = -2.196.$$

Therefore, $P = 0.029$. With the alternative version of the test, we get $t = -2.200$ (df $= 230$, $P = 0.029$). As expected, the differences between the two versions of the test are irrelevant.

Figure 2 shows the distribution of job satisfaction in Chile and Mexico. Although the departure from normality is clear here, you should not worry much about that. First, we have more than 100 observations in each group. Second, even if they can never lead to normal distributions, Likert scales typically have distributions which are not too skewed and, by definition, cannot have extreme values, which are the main concern when assuming normality in this type of tests.

¶ Source: S Poelmans & MA Canela, Statistical analysis of the results of a nine-country study of Latin managers, XIth European Congress on Work and Organizational Psychology (Lisboa, 2003).

The t test for paired data

The independence of the samples in the two-sample t test is not a trivial issue, since the distribution of the test statistic under the null can be far from a Student t if we relax the independence assumption. This typically happens in the so called **paired data**. The expression is typically used for a sample of two (potentially correlated) variables, related to the same phenomenon, like two measures taken before and after a treatment is applied.

A paired data set is not regarded as two univariate samples, but as one bivariate sample. To test the equality of means, we calculate a difference for each two-dimensional observation, testing the null $\mu = 0$ for the variable thus obtained (as in Example 1). Thus, the t test for paired data is nothing but a one-sample t test applied to the difference.

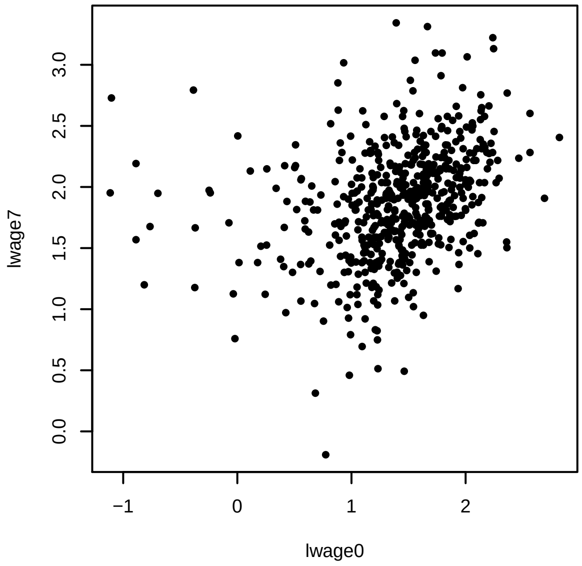

Figure 3. Correlation (Example 1)

Example 1 (continuation). The test performed in Example 1 is usually presented as paired data t test. Nevertheless, the data set can be presented as a two-sample data set, which may be confusing, because the expression “two-sample data” implicitly tells that the samples are obtained independently. But, happens if we apply a two-sample test to the `lwages` data set? We get then $t = 15.19$ (df $= 1088$, $P < 0.001$).

So, the two tests would lead to the same conclusion for this example. Is this true in general? The answer is no. In this case, in spite of the results being similar, the two-sample test can be easily shown to be wrong, because the two variables are positively correlated ($r = 0.310$), which makes sense, since most of the people with higher wages this year still get high wages seven years later. We will see later how to test the null $\rho = 0$. Figure 2 illustrates this point.

Homework

A. Reanalyze the data of Example 1 after the wages putting back in the original scale (no logs). How does the interpretation of the mean difference change?

B. Use the `scapital` data set to test the **gender effect**, that is, the mean difference between male and female employees for the three dimensions of internal social capital.