[STAT-16] Two-sample t tests

Miguel-Angel Canela
Associate Professor, IESE Business School

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, in which you can run both variants in seconds. In practice, the p-values of the two tests are very close, except (possibly) when 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 life data analysis, but, under the equal variances assumption, we have 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.

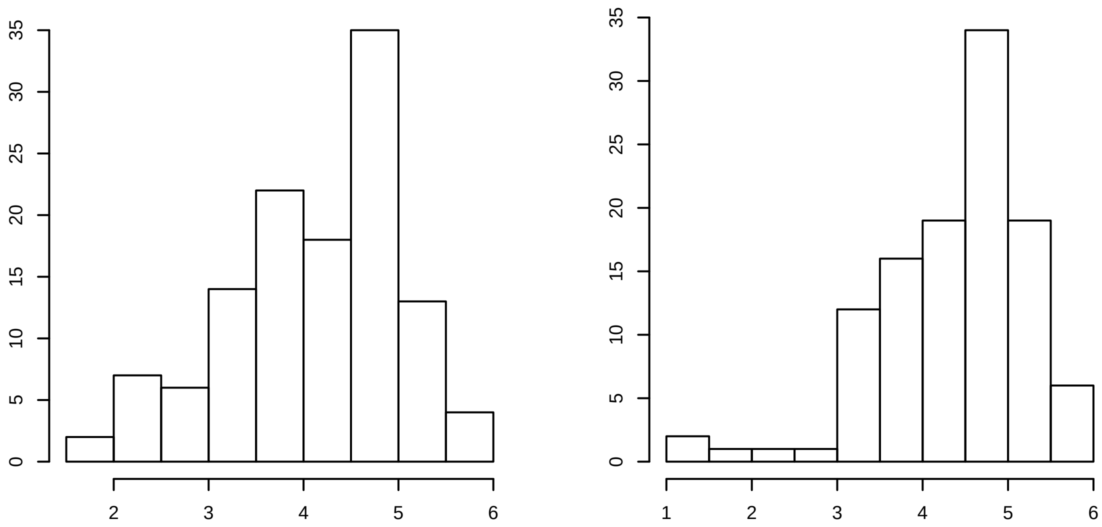

Figure 1. Distribution of job satisfaction in Chile and Mexico (Example 1)

Alternative version

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

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

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 1. 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 1)

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

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

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

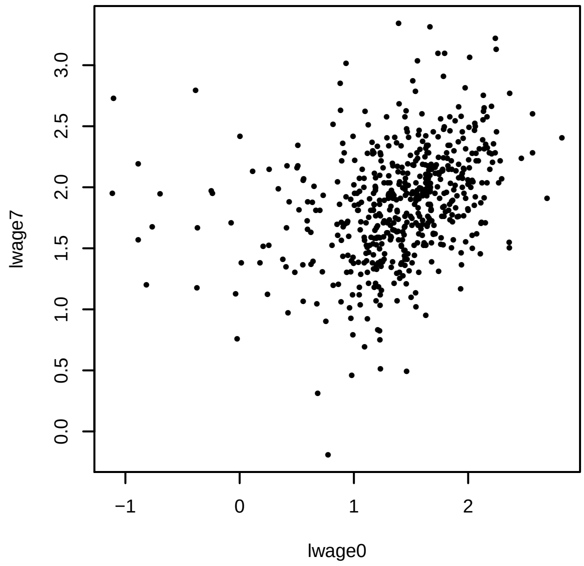

Figure 2. Correlation (Example 2)

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.

¶ 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 that 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 (see the example of lecture STAT-15). Thus, the t test for paired data is nothing but a one-sample t test applied to the difference.

Example 2. The `lwages` data set includes data on wages in years 1980 and 1987, in log scale. 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. Do these data support that there has been a change in the wages?

The analysis is usually carried out as paired data t test, getting $t = 18.22$ (df = 544, $P = 1.54 \times 10^{-58}$). Nevertheless, the data can presented as a two-sample data set. This would be wrong, because the expression "two-sample data" implicitly tells that the samples are obtained independently. But, what if we do the wrong thing, applying a two-sample test? We get then $t = 15.19$ ($\mathrm{df} = 1088$, $P < 0.001$).

So, the two tests lead to the same conclusion. 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 still get high wages seven years later. We will see later how to test the null $\rho = 0$. For the moment being, I illustrate this question with Figure 2.

Homework

A. 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.