7. The *t*-distribution, confidence intervals, and *t*-tests

The t-distribution

For any fixed value X, a *t*-value can be computed from a sample of a quantitative random variable using this formula:

$$t = \frac{X - \bar{x}}{s_{\bar{x}}}$$

where, $\bar{x}$ is the sample mean and $s_{\bar{x}}$ is its associated standard error. Recall here that $\bar{x}$ is the estimate of the population mean and $s_{\bar{x}}$ quantifies its accuracy. As a result, the ***t-value* represents the estimate of the difference between X and the population mean**. Because $\bar{x}$ is a random variable, *t*-value is also a random variable, and its probability distribution is called the ***t*-distribution**. Its shape is closely similar to Z (standard normal distribution). In contrast to Z, the *t* distribution has a single parameter – the number of degrees of freedom, which equals the number of observations in a given sample minus 1. In fact, *t* approaches Z asymptotically for high df (Fig 7.1). Similarly to the normal distribution, the *t*-distribution is symmetric, and its two tails must be considered when computing probabilities {Fig 7.2).

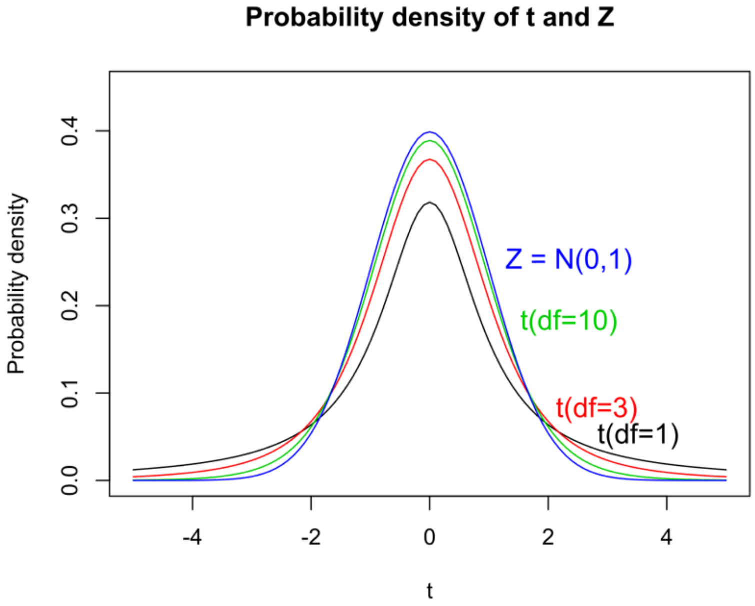

Fig. 7.1 Probability density plot of t-distributions with different DF and their comparison to standard normal distribution (Z).

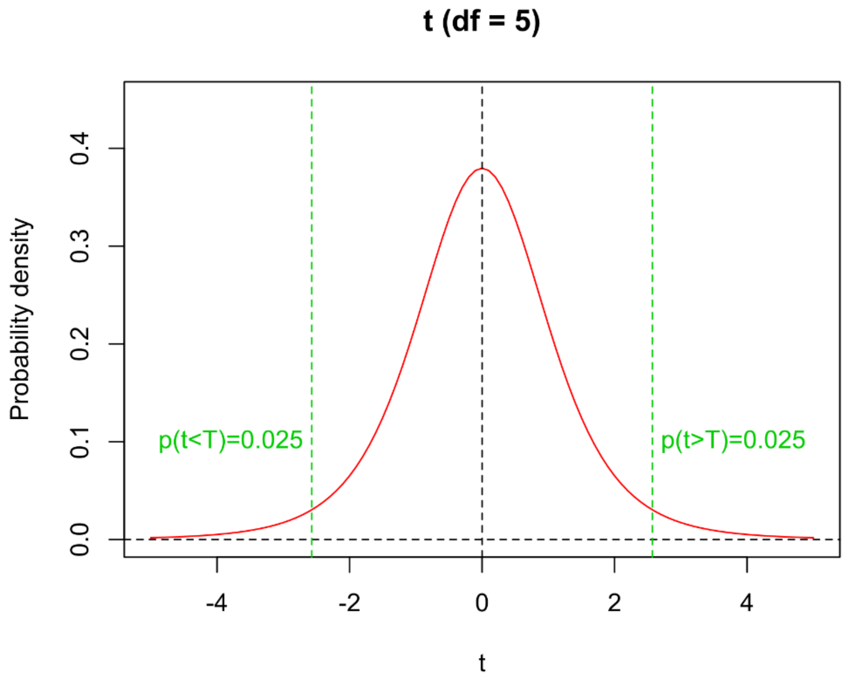

Fig 7.2. The *t*-distribution with its two tails and the 2.5% and 97.5%-quantiles.

Confidence intervals for the mean value and single sample t-test

The *t*-distribution can be used to compute confidence intervals (CI), i.e. intervals within which the population mean value lies with a certain probability (usually 95%). The confidence limits (CL) within which the CI lies are determined using these formulae:

$$CL_{low} = \bar{x} + t_{(df,p=0.025)} s_{\bar{x}}$$

$$CL_{high} = \bar{x} + t_{(df,p=0.975)} s_{\bar{x}}$$

where $t_{(\mathrm{df,\,p})}$ equals 2.5% or 97.5% probability quantile of *t*-distribution with given df. These intervals can be used as error bars in barplots or dotcharts. In fact, they represent the best option to be used like this (in contrast to standard error or 2 x standard error).

Confidence intervals can also be used to determine whether the population mean differs significantly from a given value: a value lying outside the CI is significantly different (at 5%-level of significance) while a value lying inside is not. This is closely associated with the **single sample t-test**, which tests a null hypothesis that a value X equals the population mean. Using the formula for *t*-value and DF, the *t*-test determines the probability of type I error associated with rejection of such a hypothesis.

Student t-test

If means can be compared with an *apriori* given value, two means of different samples should also be comparable. This is done by a two-sample t-test[1], which quantifies uncertainty about the values of both means considered:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are arithmetic means of the two sample and $s_{\bar{x}_1 - \bar{x}_2}$is the standard error of their difference. The $s_{\bar{x}_1 - \bar{x}_2}$ is computed using the following formula:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

where s_p^2 is the pooled variance of the two samples, and n_1 and n_2 are sample sizes of the two samples. Pooling variance like this is only possible if the two variances are equal. Correspondingly, the equality of population variances, called **homogeneity of variance,** is one of the *t*-test assumptions. In addition, the *t*-test assumes that the samples come from populations that are distributed normally. There is also the universal assumption that individual observations are independent.

The *t*-test is relatively robust to violations of the assumptions about the homogeneity of variance and normality (i.e. their moderate violation does not produce strongly biased test outcomes). If variances are not equal, Welch approximation of t-test (Welch t-test) can be used instead of the original Student *t*-test. A slightly modified formula is used for the *t*-value computation, and also the degrees of freedom are approximated (as a result, df is usually not an integer). Note here that the Welch *t*-test is used by default in R. In the original (two-sample) Student *t*-test, the DF is determined as

DF = $n_1 - 1 + n_2 - 1$

where n_1 is the size of sample 1 and n_2 is the size of sample 2.

Paired t-test

Paired *t*-test is used to analysis of data composed of paired observations. For instance, a difference in length between left and right arms of people would be analyzed using a paired *t*-test. The null hypothesis, in this case, is that the difference within the pair is zero. In fact, a paired *t*-test is fully equivalent to a single sample t-test comparing the within-pair difference distribution with zero. The number of degrees of freedom is determined as DF = n – 1 because there is just one sample (of paired values) in a paired t-test.

[1] Called also Student t-test after its inventor William Sealy Gosset (1976-1937) who used the pen name Student.

How to do in R

1. t distribution computations

functions pt and qt are available. For instance, qt(0.025, df) can be used to compute the difference between the lower confidence limit and the mean.

2. t-test

Function t.test. For two samples, the best way is to use a classifying factor and response variable in two columns. Then, t.test(response~factor) can be used. But t.test(sample1, sample2) is also okay.

important parameters:

var.equal – switches between Welch and Student variants. Defaults to FALSE (Welch)

mu – a priori null value of the difference (relevant mainly for a single sample test)

paired – TRUE specifies a paired t-test analysis.