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

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

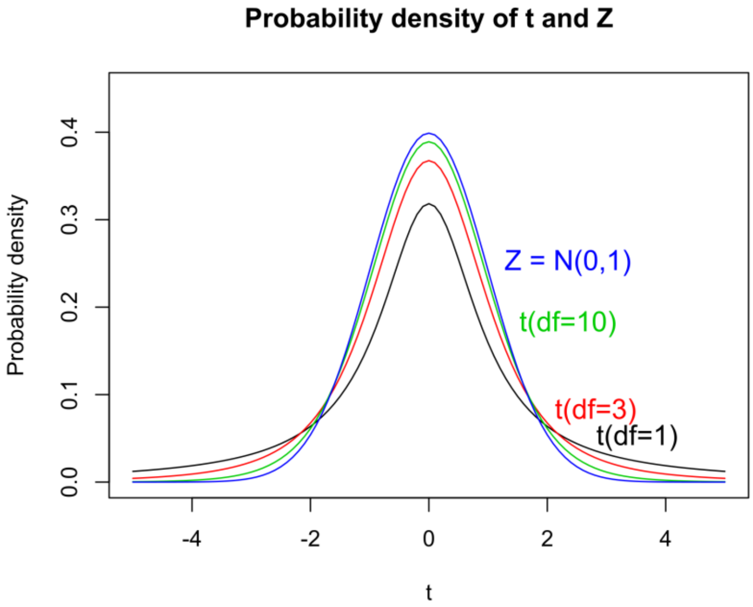

Fig. 7.1 Probability density plot of t-distributions with different DF and their comparison to standard normal distribution (Z).

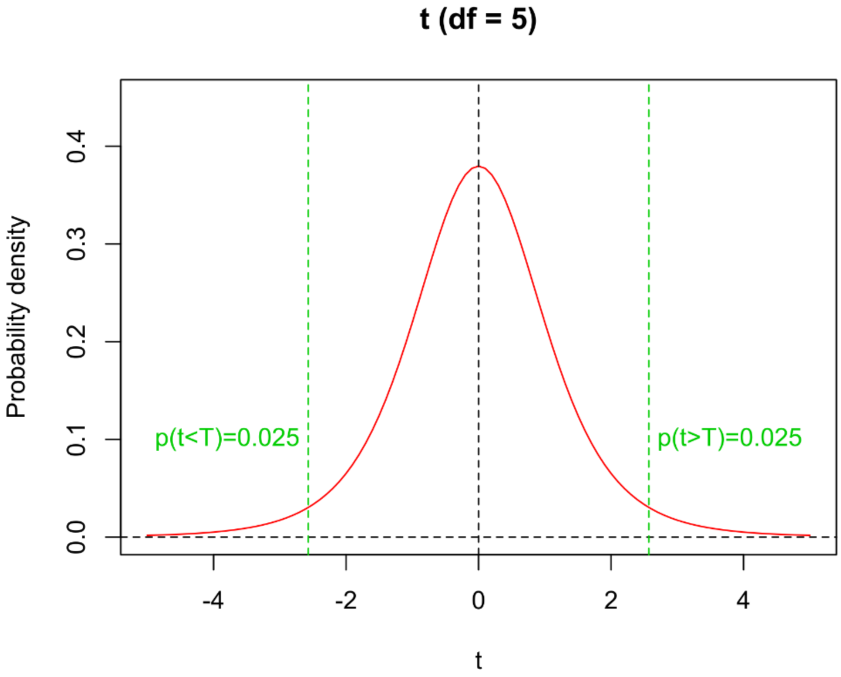

Fig 7.2. *t*-distribution with its two tails and 2.5% and 97.5%-quantiles.

Confidence intervals for mean value and single sample t-test

t-distribution can be used to compute confidence intervals (CI), i.e. intervals within which the population mean value lies with 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 $t_{(df,\ p)}$ 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 population mean is significantly different 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 **single sample t-test**, which tests a null hypothesis that given values X equals the populations mean. Using the formula for *t*-value, and DF, the t-test determines type I error probability associated with rejection of such hypothesis.

Student t-test

If means can be compared with an *a-priori* given value, two means of different samples should also be comparable with each other. This is done by 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 standard error of their difference. This is then computed using 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 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. Equality of population variances, called **homogeneity of variance** is one of the *t*-test assumptions. In addition, *t*-test assumes that the samples come from populations that are normally distributed. There is also the universal assumption that individual observations are independent.

t-test is relatively robust to violations of the assumptions about 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 *t*-value computation and also the degrees of freedom are approximated (as a result, DF is usually not an integer). Note, that Welch *t*-test is used by default in R. In original (two-sample) Student *t*-test, the DF is determined as

DF = $n_1 - 1 + n_2 - 1$

Paired t-test

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

[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 lower confidence limit and the mean.

2. t-test

Function t.test. For two sample, 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 for single sample test)

paired – TRUE specifies a paired t-test analysis.