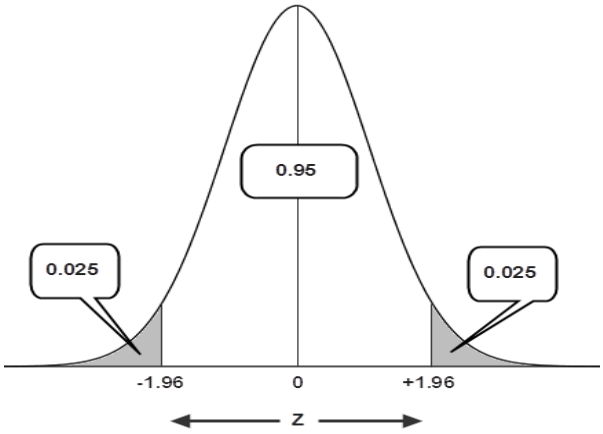
Overcoming the limitations of t-distribution

We know well that if the population of data are normally distributed, the sampling distributions of the means and the difference between the means are also normal. Even if the parent populations are not normally distributed, the Central Limit Theorem provides that the sampling distributions will approach normality when the sample size is sufficiently large. The theorem gives the following equation:

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} = \bar{x} \pm z\sigma_M$$

where, μ is the population mean; $\bar{x}$, mean of sample means; $\sigma_M = \sigma/\sqrt{n}$, standard error of mean and z, standard normal variable which has a mean of zero and a standard deviation of 1.

For the two-tailed case, the probability of a mean value falling outside the range $\mu \pm 1.96\sigma/\sqrt{n}$ is 0.05, meaning z = 1.96 with 95% confidence:



However, it may be noted that the statistics $\bar{x}$ and s are estimates of the unknown parameters μ and σ, and usually approach them more closely when n increases. But $\bar{x}$ and s are variables which if a series of measurements were repeated, the resultant values of $\bar{x}$ and s would be different each time. Hence, we cannot take them directly to substitute for the parameters μ and σ in the above equation.

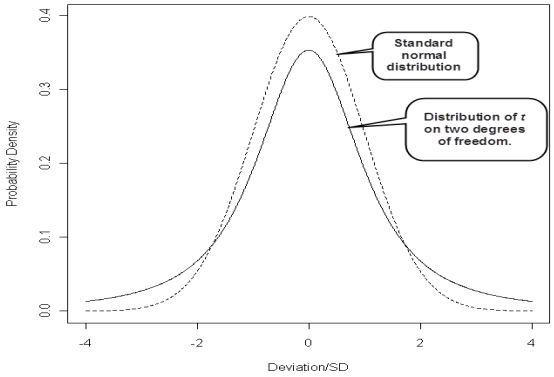
Instead, we can use modified equations in which we substitute a variable t (also called Student's t) for the standard normal variable, z as below:

$$\mu = \bar{x} \pm t \frac{s}{\sqrt{n}} = \bar{x} \pm s_M$$

When n approaches infinity, t approaches z. Hence, we can re-write as:

$$t = \frac{(\mu - \bar{x})}{s_M}$$

The figure below shows a comparison of a t-distribution of 2 degrees of freedom which has 'thicker' tails, with the standard normal distribution:



When we want to study if the means of two independent samples A and B upon experiment, drawn from a population of which its population standard deviation σ is unknown, we use the following t-distribution equation:

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \quad \text{where pooled} \quad s_p = \sqrt{\frac{(n_A - 1)s_A^{\,2} + (n_B - 1)s_B^{\,2}}{(n_A - 1) + (n_B - 1)}}$$

There is an important assumption for the use of the above equation. That is, the pooled standard deviation s_p is an estimate of the supposedly constant population variance σ^2. Hence the t-test has assumed homogeneity of variance under normal distribution.

The t-test has been said to be robust to some violation of the assumptions of the underlying statistical model. Indeed, computer simulations have shown that even with moderate violations of the above assumptions, the error rates are little affected, provided the sample sizes are not too small with no data outliers and the samples are of equal or nearly equal size.

But there are limits to this robustness. When the sample variances are very different, especially in combination with markedly discrepant sample sizes, error rates can become unacceptably high, say more than 5%, threatening the assumption of homogeneity. When encountering this situation, we can carry out the Behrens-Fisher T-statistic test which uses the following equation:

$$T = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)}}$$

The statistic T is distributed approximately as *t*, but on fewer degrees of freedom than $n_A + n_B - 2$. Several formulae for the degrees of freedom of T have been proposed and the most popular one is the Welch–Satterthwaite's:

$$df_T = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1}}$$

The above equation shows that the greater the disparity between the two sample variance estimates, the smaller will be the df_T value and this will cause a greater T for the test to show significance. There will be situations in which T may be significant when t is not.