# *t model*

*Date: 2017-10-31*

## *What it is*

The $t$ model is a modification of the normal model which is often used in practice when dealing with observed data. Like the standard normal model $Z$, $t$, is symmetric around 0. Compared to $Z$, $t$ models have "heavier tails," meaning there is more probability in the areas away from 0.

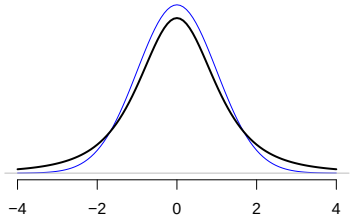

Figure 1: A $t$ model with df=3 (black curve) compared to a Z model (blue curve).

## *When to use it*

The $t$ model is most often used when performing statistical inference (confidence interval and hypothesis tests) using a statistics which is theoretically normally distributed. We use $t$ instead of a normal model when we do not know the values of the parameters, especially the population standard deviation, $\sigma$. We see the $t$ model most often when performing inference for (a) a single mean, (b) the difference between two means (independent or paired samples), and (c) linear regression coefficients.

## *How to use it*

The $t$ model has a parameter called the *degrees of freedom* ($df$). Calculations for $df$ are usually pretty simple, but they depend on the context of the problem. The table below gives the formulas for calculating $df$ in the cases outlined above:

| Case | Parameter | $df$ |
|---|---|---|
| (a) | Single mean: $\mu$ | $n-1$ |
| (b-1) | Difference between two means, independent samples: $\mu_1 - \mu_2$ | $\min\{n_1, n_2\} - 1^{(1)}$ |
| (b-2) | Difference between two mean, paired samples: $\mu_d$ | $n_d - 1^{(2)}$ |
| (c) | Linear regression coefficient: $(\beta_j)$ | $n - p - 1^{(3)}$ |

In the table, (1) $n_1$ and $n_2$ are the sizes of samples 1 and 2, respectively; (2) $n_d$ is the number of paired differences; and (3) $p$ is the number of predictors in the regression model.

To get critical values in R, use the function `qt()`. For example, suppose you have a sample size of 50 observations and want to compute a 95% confidence interval for $\mu$. The correct critical $t$ value would be[1]

```
qt(0.975, df = 49)
```

[1] To obtain the middle $1-\alpha$ of a distribution, you want $\pm$ the $1-\alpha/2$ quantile.

```
## [1] 2.009575
```

Or, say you have two samples, one of size 70 and the other of size 90, and you want to compare the two means. A 90% confidence interval would use the critical $t$[2]

```
qt(0.95, df = 69)
```

```
## [1] 1.667239
```

[2] To get $df$, take the smaller sample size, 70, and subtract 1.

## *Why it works*

For simplicity, let's stick with to the context of estimating a single population mean. The Central Limit Theorem says that for a large sample, $\bar{y}$ is approximately normally distributed and that

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim Z$$

However, $\sigma$ is the *population* standard deviation, which we do not know. We can replace $\sigma$ with its estimate, $s$. The new statistic is no longer a $Z$ distribution, but a $t$ distribution:[3]

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t^{(df)}$$

This substitution of $s$ for $\sigma$ adds extra uncertainty and variability to our statistic, leading to the heavier tails described earlier.

The degrees of freedom for the $t$ model are a measure of the accuracy of $s$ in estimating $\sigma$. The larger the degrees of freedom, the better estimate $s$ is of $\sigma$ and the less uncertainty there is in that substitution. This means that $t$ models with larger $df$ look much more like a normal model. In fact, $t \to Z$ as $df \to \infty$:

[3] To really get into the weeds, $t$ is defined theoretically as $t = \frac{Z}{\sqrt{X/df}}$, where $Z$ is a standard normal distribution and $X$ is a chi-squared distribution with $df$ degrees of freedom (It turns out that $(n-1)s^2/\sigma$ has a chi-squared distribution with $n-1$ degrees of freedom).

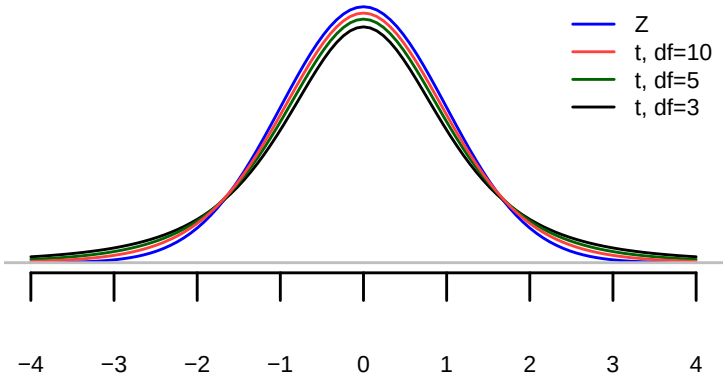


Figure 2: The $t$ model approaches $Z$ as degrees of freedom get larger.