# Calculate the confidence interval using student's t distribution



**1 author:**

Chuanli Jiang
NWRA - NorthWest Research Associates

# Calculate the confidence interval using student's $t$ distribution

**ChuanLi Jiang**[*]

A frequently asked scientific question about a time series is whether there is a linear trend in the variable $y(x)$, that is, can we describe the time series as $y(x) = b_1 \cdot x + b_0 + \epsilon$? If we then perform a linear regression ($\hat{y(x)} = \hat{b_1} \cdot x + \hat{b_0}$) to obtain the trend $b_1$, how do we know if the trend is statistically significant?

The first step is to calculate the confidence interval (for example, 95%) for the true slope of the linear trend. If both the lower and upper limits of the interval are larger (smaller) than zero, you could conclude that the assumed linear trend is increasing (decreasing). Section 3 will discribe how to compute the confidence interval for the true slope of a linear trend.

As a second check we could calculate the 95% confidence interval for the true mean. If the trend falls outside of this interval, we would have additional confidence that there is a linear trend. Section 2 will discribe how to compute the confidence interval for the true mean.

## 1. student's $t$ distribution

Let X be a standard normal random variable, $X \sim N(0, 1)$, and let Y be a chi-square random variable with $\nu$ degrees of freedom, $Y \sim \chi^2(\nu)$. Then if $X$ and $Y$ are independent, the student's $t$ distribution can be derived as follows (*Mendenhall, Wackerly and Scheaffer* 1990):

$$t(\nu) = \frac{X}{\sqrt{Y/\nu}} \tag{1}$$

For example, $Z = \sqrt{N^*}(\overline{y(x)} - \mu)/\sigma^2 \sim N(0, 1)$, and $(N^* - 1)s^2/\sigma^2 \sim \chi^2(\nu)$ (*Mendenhall, Wackerly and Scheaffer* 1990). Therefore,

$$t(\nu) = \frac{\overline{y(x)} - \mu}{\frac{s}{\sqrt{\nu}}} \tag{2}$$

follows the student's $t$ distribution. Here $\mu$ is the true mean, $s$ is the sample standard deviation, and $\nu = N^* - 1$ is the degrees of freedom (use of the mean reduces the degrees of freedom by one; $N^*$ is the effective degrees of freedom, it is different from

[*]chuanlij@ocean.washington.edu

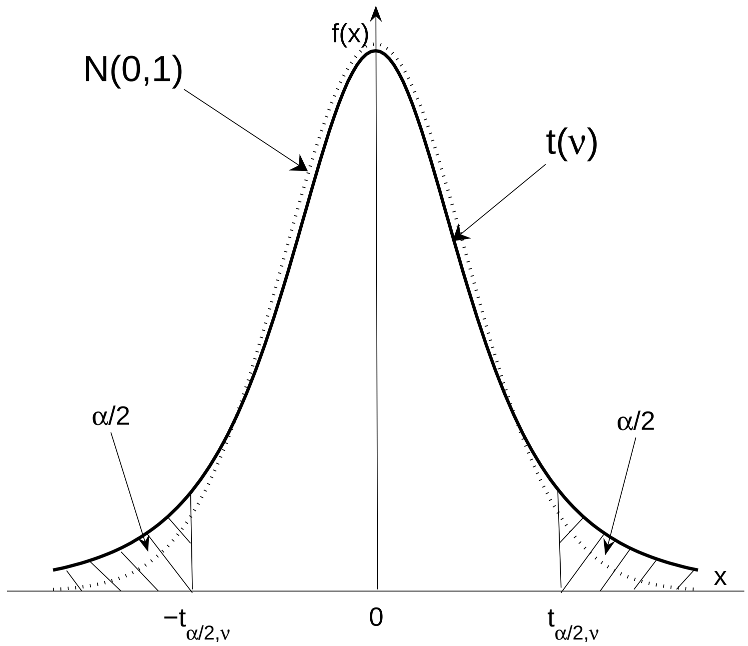


Figure 1: Student's $t$ distribution

the size of the sample). Student's $t$ distribution is usually used for a small sample. It approaches the normal distribution for larger $\nu$ (see Fig.1). Note that the $t$ distribution is symmetric.

Values of $t$ for specified significance levels ($1-\alpha$) and degrees of freedom ($\nu$) can be obtained from tables (one-tailed table in *Hartmann* (2007b) or two-tailed Table $D$.3b in *Emery and Thomson* (2001)). Here we are going to use one-tailed table only. For a specified significance level 95% ($\alpha = 0.05$), the probability of $|t| \leq t_{\alpha/2}(\nu) = t_{0.025}(\nu)$ (the corresponding area under the probability density function $f(x)$ as shown in Fig. 1) is $95\%$:

$$P(|t| \leq t_{\alpha/2}(\nu)) = 1 - \alpha \quad (3)$$

Here $t_{\alpha/2}(\nu) = t_{0.025}(\nu)$ is the value of $t$ for which only 5% would be expected to be greater (right-hand tail table).

## 2. 95% confidence interval for the true mean $\mu$

There is 95% probability that the $t$ distribution falls in the following interval (can be easily derived from Equation. 2 and 3):

$$t_{-0.025} \leq \frac{\overline{y(x)} - \mu}{s} \cdot \sqrt{\nu} \leq t_{0.025}$$

Therefore, the true mean $\mu$ is expected with 95% confidence to lie in the interval:

$$\overline{y(x)} - t_{0.025} \cdot \frac{s}{\sqrt{\nu}} \leq \mu \leq \overline{y(x)} + t_{0.025} \cdot \frac{s}{\sqrt{\nu}} \quad (4)$$

where, degrees of freedom $\nu = N^* - 1$. Note that $N^*$ is the effective degrees of freedom, it is not the size of the samples. Detailed information can be found in *Hartmann* (2007a).

# 3. 95% confidence interval for the true slope of a linear trend $b_1$

Because linear regression coefficients $b_1$ and $b_0$ are linear functions of random variables $x_i$, they will have normal distributions if the $x_i$ have normal distributions. Then,

$$\hat{b_1} \sim N(b_1, \frac{\sigma^2}{\sum(x_i - \overline{x})^2})$$

If let $l_{xx} = \sum(x_i - \overline{x})^2 = (N-1) \cdot s^2$ ($N$ is the size of the samples, and s is the sample standard deviation), then

$$Z = \frac{\hat{b_1} - b_1}{\frac{\sigma}{[l_{xx}]^{1/2}}} \sim N(0, 1)$$

We also know that:

$$\frac{S_E}{\sigma^2} \sim \chi^2(\nu)$$

where, $\nu = N^* - 2$ is the degrees of freedom because two parameters are needed for any linear regression estimate, $S_E = \sum (y_i - \hat{y_i})^2$ is the sum of the squared errors (the total variance that is not explained by the linear regression).

Then the student's $t$ distribution can be written as:

$$t = \frac{\frac{\frac{\hat{b_1} - b_1}{\sigma}}{[l_{xx}]^{1/2}}}{\sqrt{\frac{S_E}{\sigma^2 \cdot \nu}}} = \frac{(\hat{b_1} - b_1) \cdot [l_{xx}]^{1/2}}{\sqrt{\frac{S_E}{\nu}}} = \frac{(\hat{b_1} - b_1) \cdot [l_{xx}]^{1/2}}{S_\epsilon} = \frac{\hat{b_1} - b_1}{\frac{S_\epsilon}{[l_{xx}]^{1/2}}} \sim t(\nu)$$

Where, $S_\epsilon = \sqrt{S_E/\nu}$ is the standard error of the linear regression line estimate.

Therefore, the 95% confidence interval for the true slope $b_1$ is:

$$\hat{b1} - t_{0.025} \cdot \frac{S_\epsilon}{[l_{xx}]^{1/2}} \leq b1 \leq \hat{b1} + t_{0.025} \cdot \frac{S_\epsilon}{[l_{xx}]^{1/2}} \quad (5)$$

Some information can also be found in *Bendat and Piersol* (2000).

# References


Emery, W. J., and R. E. Thomson, *Data analysis methods in physical oceanography*, ISBN 0444507574, Elsevier, 2001, P238.

Hartmann, D. L., *ATM552 Notes: Statistics Review*, `http://www.atmos.washington.edu/~dennis/552_Notes_ftp.html`, 2007.

Hartmann, D. L., *ATM552 Notes: Tables Chap1*, `http://www.atmos.washington.edu/~dennis/552_Notes_ftp.html`, 2007.

Mendenhall, W., D. D. Wackerly., and R. L. Scheaffer, *Mathematical Statistics with Applications*, ISBN 0534920268, PWS-Kent, Boston,1990, P308, P311.

Bendat, J. S., and A. G. Piersol, *Random Data Analysis and Measurement Procedures*, ISBN 0471317330, 2000, P113.