Notes

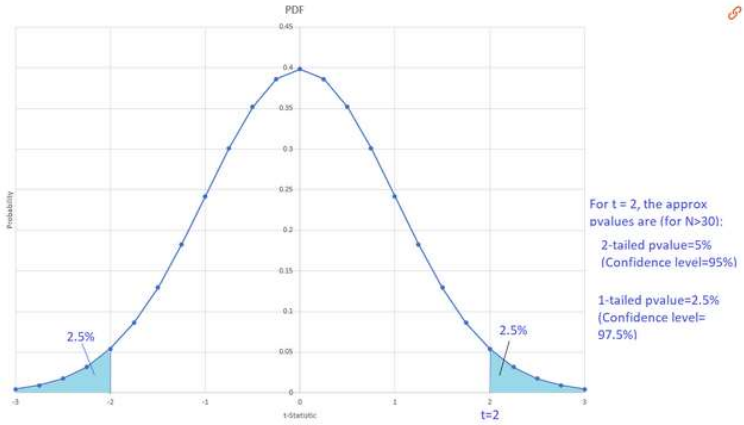
Monday, August 15, 2022 8:56 AM

Myroslava Sánchez Andrade A01730712 | 15/08/2022

3 Confidence level, Type I Error and p-value

The **confidence level** of a test is related to the error level of a test.
For a 95% confidence level, we can end up in a mistaken conclusion 5% of the time. This error is also called the **Type I Error**.

The **p-value** is the probability that we will be wrong if we reject the null hypothesis.



When the number of observations of the sample increases, the t-Student distribution approximates the Z (mean = 0, std = 1) normal distribution.
The 2-tailed p-value will be twice the value of the 1-tailed p-value since the t-Student distribution is symmetric.
We always want to have a very small p-value in order to reject H0. The 2-tailed p-value is a more conservative value.
We can define the p-value of a t-test (in terms of confidence level of test) as:

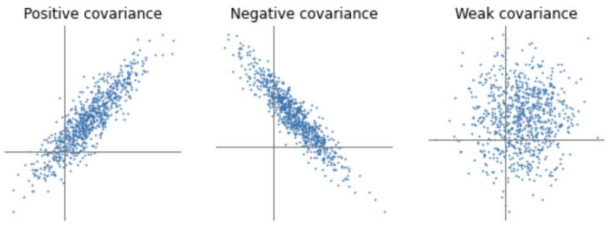
$$\text{pvalue} = (1 - \text{ConfidenceLevel})$$

4 Measures of linear relationship

The linear (proportional change) relationships measure whether there is a pattern of *movement* for a random variable when another variable moves up or down. The main two measures of linear relationship between 2 random variables are:

- Covariance

Is a measure of the joint probability of two random variables. If the greater the values of one variable mainly correspond with the greater values of the other variable => the covariance is positive [similar behavior]; if the greater values of one variable mainly correspond to the lesser values of other => the covariance is negative [opposite behavior].
The **sign** of the variance shows the tendency in the linear relationship between the variables. We cannot understand the magnitude of covariance; only its sign (+, -, 0).



The Covariance is the average of product deviations between X and Y from their corresponding means.

$$\text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Sample covariance

$$\text{Cov}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Correlation

Statistical relationship between two random variables. It refers to the degree to which a pair of variables are linearly related, thus, the interpretation of the Covariance in percentage. The result will have values between -1 and 1+.

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{SD(X)SD(Y)}$$

Then:

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where, $\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

If Corr(x, y) = 0.30, then about 30% of the cases, when x goes up, y goes up; and when x goes down, y goes dow.
If Corr(x, y) = -0.30, then about 30% of the cases, when x goes up, y goes down, and when x goes down, y goes up.

If we want to test that Corr(X,Y) is **positive and significant**, we need to do a hypothesis test. The formula for the standard error (standard deviation of the correlation) is:

$$SD(corr) = \sqrt{\frac{(1 - corr^2)}{(N - 2)}}$$

Then, the t-Statistic for this hypothesis test will be:

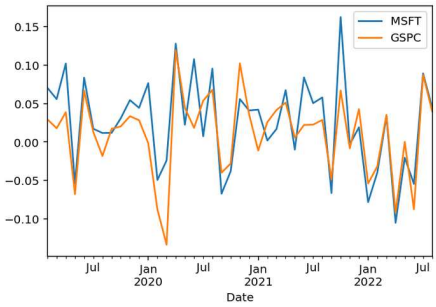
$$t = \frac{corr}{\sqrt{\frac{(1-corr^2)}{(N-2)}}}$$

Regresiones espurias

En el modelo de regresión lineal ocurre cuando la variable independiente contribuye a explicar la variabilidad de la variable respuesta, a pesar de que evidentemente las variables no tienen relación. Se necesita de pruebas de co-integración para determinar si existe una verdadera relación entre las variables.

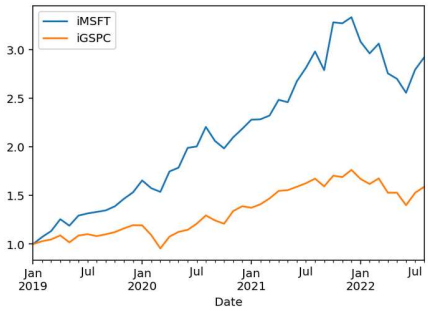
Stationary variables

Variables that have a growing or declining trend over time. This variables have a similar average and standard deviation in any time period. (**Stock returns** behave like stationary variables).



Non-stationary variables

Variables that usually grow over time (sooner or later), this variables change its mean depending on the time period. (**Stock prices** usually grow over time).



In statistics we have to be very careful when looking at linear relationships when using non-stationary variables. It is very likely to end up with **spurious** measures of linear relationships.

Regresión lineal => predicción para regresión (modelado de una variable)
Regresión logísitca => predicción para clasificación