Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Goodness of fit

For measure the **goodness of fit** we use the **chi-square test**, that is to say the same test to measure the independency of variables.

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

O_i : Observed i value.
E_i : Expected i value.

$$\boxed{E_i = F(Y_i^{upper}) - F(Y_i^{lower})}$$

$F(\cdot)$: The cumulative distribution function for the probability distribution being tested.
Y_i^{upper} : The upper limit for the i value.
Y_i^{lower} : The lower limit for the i value.
n : Sample size.

The **chi-square** distribution has $(k - c)$ degrees of freedom whe k is the number of non-empty cells and c is the number of estimated parameters for the distribution **plus one.**

Example: For independence of variables we have $(k - c) = (N_{rows} - 1) \cdot (N_{columns} - 1)$ degrees of freedom.

$$(k - c) = (N_{rows} - 1) \cdot (N_{columns} - 1) = N_{rows} \cdot N_{columns} - N_{rows} - N_{columns} + 1$$

$$= (N_{rows} \cdot N_{columns}) - (N_{rows} + N_{columns} - 1)$$

$$= (N_{rows} \cdot N_{columns}) - ([N_{rows} - 1] + [N_{columns} - 1] + 1)$$

We can see that:

$$\boxed{k = N_{rows} \cdot N_{columns}}$$

The estimated parameters for the columns are:

$$\boxed{c_{columns} = N_{columns} - 1}$$

because the probability have to sum 1 and we have to estimate one less parameter.

The estimated parameters for the rows are:

$$\boxed{c_{rows} = N_{rows} - 1}$$

because the probability have to sum 1 and we have to estimate one less parameter.

And for this, the value of c is:

$$\boxed{c = c_{row} + c_{columns} + 1}$$
