

Symmetric Uncertainty is related to Chi-Squared

SOLABIMA 2026 - Paraguay

Santiago Gómez Guerrero, DSc (Facultad Politécnica)
Christian Emilio Schaerer, DSc (Facultad Politécnica)
Inocencio Esteban Ortiz Samudio, DSc (Facultad de Ingeniería)
Universidad Nacional de Asunción, Paraguay

The beginning of correlation: Pearson and Spearman

- ▶ During the 20th century, statistical correlation was understood as the possible association between the values of two variables. The basis was the pioneering work of Pearson [1] in 1895, stating correlation as:

$$r(X,Y) = \text{cov}(X,Y) / (\sigma_X \sigma_Y)$$

- ▶ Spearman's ρ (rho) ([2] 1904) was the same as Pearson's correlation, except that it was applied to the ranked values of X and Y .
- ▶ Thus, both proposals required that the two variables considered be quantitative integer or real variables. For qualitative variables, covariances and standard deviations could not be calculated, so it was not possible to compute r and ρ in these cases.

Hall's proposal

- ▶ In 1998, Hall [3] proposed a correlation formula that overcomes this impediment, based on entropies:

$$SU(X, Y) := 2 \left[1 - \frac{H(X, Y)}{H(X) + H(Y)} \right]$$

where $H(X)$ is the entropy of a random variable X .

- ▶ From Quinlan ([6] 1993) and other results from conditional entropy and mutual information, Hall proved that this formula featuring the quotient between $H(X, Y)$ and $H(X) + H(Y)$ yields in fact a correlation between X and Y .
- ▶ Entropy has roots in physics (thermodynamics) and was later applied to information theory. In our case, we found a way to interpret it using probabilities.

Understanding entropy

- From a histogram of X , think about the surprise of a result: for an observer, a “rare” category causes greater surprise than a relatively “familiar” one, and we can define surprise as

$$\pi := 1/P(X).$$

- If we invert $0 \leq P(X) \leq 1$ term by term we have $1 \leq \pi \leq \infty$. Seeking to avoid having to deal with large values of π , it is handy to use the logarithmic function’s slow and monotonic growth in $[1, \infty)$; thus we take $\log(\pi)$ instead of π . Thus, for a random variable X with categories $x_1, \dots, x_n$, the expected value of $\log(\pi)$ is

$$E(\log(\pi)) = E\left(\log \frac{1}{P(X)}\right) = \sum_{i=1}^n P(x_i) \log \frac{1}{P(x_i)} = - \sum_{i=1}^n P(x_i) \log(P(x_i))$$

- The last formula matches Shannon's definition of entropy. This entropy evaluates how surprising the variability of X is. Note then that in calculating the entropy of a qualitative variable, we only use the probability of each category. Categories with 0 probability are not considered part of X .

Defining the proposed measure

- From Hall's SU proposal for two variables, we extend the concept of correlation to n variables. For multiple qualitative variables, the newly proposed **Multivariate SU** (or MSU) measure essentially includes the joint entropy of the n variables divided by the sum of their n individual entropies,

$$MSU(X_{1:n}) := \frac{n}{n-1} \left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^n H(X_i)} \right].$$

- It is shown in [4] that this ratio has values in the range $[0, 1]$. This makes it convenient and, above all, comparable with other correlation measures. Because of its component terms, MSU could also be called Multivariate Entropic Correlation.

Mathematical properties

To this date our team has shown that:

- ▶ Under MSU, zero correlation is synonymous with *independence of variables*, and vice versa (2023).
- ▶ *Interaction* within a subset of variables is measurable in a finite number of steps. We propose a formal definition of interaction, computable for a subset of variables (2022), which, together with the concept of cooperation between attributes, provides new techniques for attribute selection (2024).
- ▶ In two dimensions, SU provides a *distance measure* between n random variables taken in pairs (2025). This facilitates the mathematical treatment of relationships between abstract and often times not measurable concepts. Examples:
 - ▶ Medicine • How related is coronary disease to blood type, race and age?
 - ▶ Biology • Are forest fires related to land use, topography, dominant species?
 - ▶ Social sciences • Is employability more related to responsibility or to creativity?

Is SU related to chi-squared?

Contingency tables are conveniently used to show the relationship between two variables using observed data.

To establish whether these variables are independent or not, Pearson also developed a test. It requires to find the *chi-squared* value

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

from the contingency table, where for each cell O is the observed cell count, and the expected count E is computed as (row total x column total)/(grand total). When χ^2 is 0 or close to 0 (a threshold given by table), the hypothesis of *independence* is accepted; otherwise the variables are deemed as *not-independent*.

A contingency table configuration arises from degrees of freedom, marginal totals and observed cell counts. Each distinct contingency table configuration produces a specific χ^2 value and a specific SU value. Being originated by the same contingency table, one can think that these two values may be related.

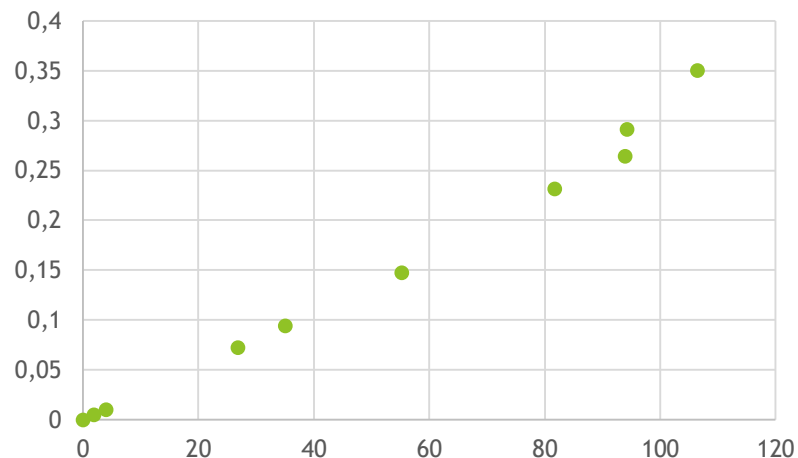
Is SU related to chi-squared? (cont.)

We ran experiments where both χ^2 and SU were computed for a variety of observed cell counts within a configuration, and then we plotted SU as a function of χ^2 . In the following, the results for two of many possible configurations are shown.

There is a clear cuasi-linear tendency in the plot, that is, knowing one of SU or χ^2 would allow for estimation of the other value.

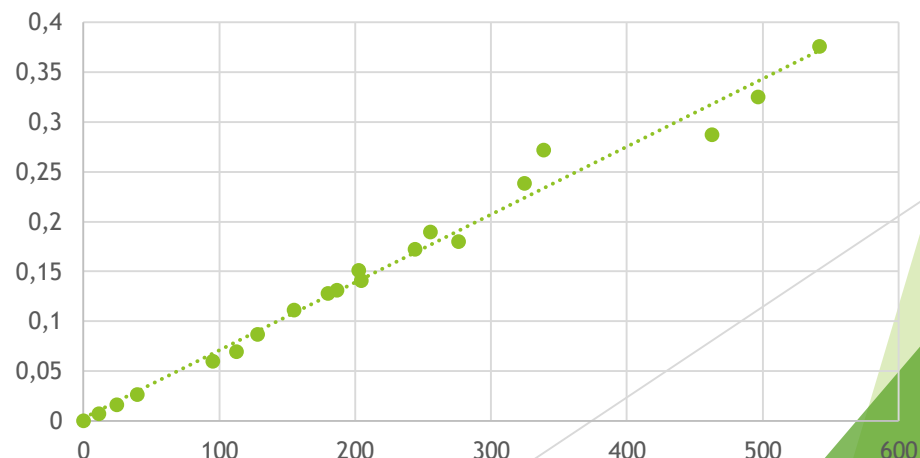
Is preference for a style of coffee dependent on the person's sex?

SU vs Chi-squared, with df=2



Is each person's neighborhood of residence independent of the person's occupational classification?

SU vs Chi-squared, with df=6



Mathematical properties - 2

We have also proven that:

- **Topologically**, the distances measured through MSU are continuous (2025). This concept arose from numerous database simulations. As an example, see a snapshot of the XOR data behavior or pattern:

3-way ABC					
A	B	$C = A \oplus B$	X	P(X)	$P(X) \log(P(X))$
0	0	0	000	0,25	-0,5
0	1	1	011	1,00E-01	-0,33219281
1	0	1	101	2,00E-01	-0,46438562
1	1	0	110	0,45	-0,51840139
Entropy					1,81497982
MSU					0,53050556

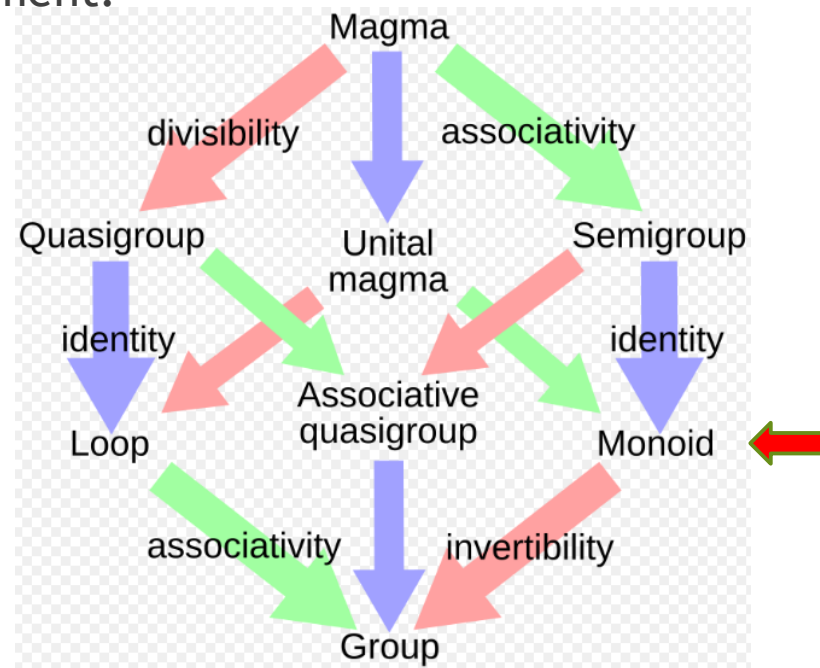
1-way A	1-way B	1-way C
$P(X) \log(P(X))$	$P(X) \log(P(X))$	$P(X) \log(P(X))$
-0,53010061	-0,47437306	-0,52108968
-0,40396744	-0,51840139	-0,36020122
0,934068055	0,992774454	0,881290899

When large samples are taken, $P(X)$ can deviate from the uniformity that we are used to. Continuity arises from the fact that each $P(X)$ may take on very many values, producing an infinite variety of histograms.

Mathematical properties - 3

In addition, we have proven that:

- **Algebraically**, a set of categorical random variables is a *commutative monoid* (2025). A monoid is a magma where an internal operation has associativity and has an identity element.



Some applications / Future work

- ▶ An MSU-based *criterion for building a decision node* was tried in a Random Forest model [5]. The new algorithm outperforms the classification accuracy of other existing Random Forest algorithms, which use the more traditional criteria (2024).
- ▶ By facilitating the *detection of interactions*, we are better able to make accurate attribute selections that represent the synergy of two or more variables.
- ▶ The distances between pairs of variables allow us to construct *clusters of attributes* (as opposed to clusters of observations). This is a way to see associations between our data, resembling some multivariate graphing methods. Variables exhibited close to each other have greater correlation.
- ▶ Most likely, the nearly linear relationship between chi-squared and SU extends to 3, 4 and more dimensions. Thus, extended versions of χ^2 and/or similar measures will have to be contrasted with SU for more general conclusions.

Is statistics better with MSU?

- ▶ The main contribution of MSU to statistical science is the formal calculation of the correlation of categorical variables in many dimensions. This calculation can be extended to any combination of multiple categorical and numerical variables, adopting discretization methods for the latter.
- ▶ MSU is based on the calculation of entropies found in the sample, without assuming an underlying distribution, and is therefore nonparametric. When reduced to two discretized numerical variables, MSU detects linear as well as nonlinear correlation.
- ▶ From a more abstract perspective, we find that statistical correlation is equivalent to the mathematical concept of *similarity* between random variables [4].
- ▶ Statistical science continues to develop measures and tools to understand the world around us. Measures that possess good mathematical properties will undoubtedly be more applicable to real world contexts.

References

- [1] Pearson, K. (1895). *Note on regression and inheritance in the case of two parents*. In Proceedings of the Royal Society of London, volumen 58, pages 240-242.
- [2] Spearman, C. (1904). *The proof and measurement of association between two things*. American Journal of Psychology. 15(1):72-101.
- [3] Hall, M.A. (1998). *Correlation-based feature subset selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.
- [4] Gómez-Guerrero, Santiago. *A Proposed Correlation Measure for Categorical Random Variables*. Doctoral thesis. Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo, Paraguay.
- [5] García-Torres, M. et al (2025). *RF_MSU: A Multivariate Symmetrical Uncertainty based Random Forest*. Pattern Recognition, volume 169, January 2026.
- [6] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.