planetmath.org

Math for the people, by the people.

Cramer's V

Canonical name	CramersV
Date of creation	2013-03-22 15:10:11
Last modified on	2013-03-22 15:10:11
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	7
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62H17
Synonym	Cramér's V
Synonym	phi coefficient
Defines	phi statistic

Cramer's V is a statistic measuring the strength of association or dependency between two (nominal) categorical variables in a contingency table.

Setup. Suppose X and Y are two categorical variables that are to be analyzed in a some experimental or observational data with the following information:

- X has M distinct categories or classes, labeled $X_1, \ldots, X_M$,
- Y has N distinct categories, labeled $Y_1, \ldots, Y_N$,
- n pairs of observations (x_k, y_k) are taken, where x_i belongs to one of the M categories in X and y_i belongs to one of the N categories in Y.

Form a $M \times N$ contingency table such that Cell (i, j) contains the count n_{ij} of occurrences of Category X_i in X and Category Y_j in Y:

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_N
X_1	n_{11}	n_{12}	$\cdots$	n_{1N}
X_2	n_{21}	n_{22}	$\cdots$	n_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_M	n_{M1}	n_{M2}	$\cdots$	n_{MN}

Note that $n = \sum n_{ij}$.

Definition. Suppose that the null hypothesis is that X and Y are independent random variables. Based on the table and the null hypothesis, the chi-squared statistic χ^2 can be computed. Then, *Cramer's V* is defined to be

$$V = V(X, Y) = \sqrt{\frac{\chi^2}{n \min(M-1, N-1)}}.$$

Of course, in order for V to make sense, each categorical variable must have at least 2 categories.

Remarks.

1. $0 \leq V \leq 1$. The closer V is to 0, the smaller the association between the categorical variables X and Y. On the other hand, V being close to 1 is an indication of a strong association between X and Y. If $X = Y$, then $V(X, Y) = 1$.

2. When comparing more than two categorical variables, it is customary to set up a square matrix, where cell (i, j) represents the Cramer's V between the ith variable and the jth variable. If there are n variables, there are $\frac{n(n-1)}{2}$ Cramer's V's to calculate, since, for any discrete random variables X and Y, $V(X, X) = 1$ and $V(X, Y) = V(Y, X)$. Consequently, this matrix is symmetric.

3. If one of the categorical variables is dichotomous, (either M or $N = 2$), Cramer's V is equal to the *phi statistic* (Φ), which is defined to be

$$\Phi = \sqrt{\frac{\chi^2}{n}}.$$

4. Cramer's V is named after the Swedish mathematician and statistician Harald Cramér, who sought to make statistics mathematically rigorous, much like Kolmogorov's axiomatization of probability theory. Cramér also made contributions to number theory, probability theory, and actuarial mathematics widely used by the insurance industry.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley-Interscience, 2nd ed. 2002.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1999.