On the maximum of Cramér's V

Etsuo Hamada

Faculty of Information Science, Osaka Institute of Technology,
1-79-1 Kitayama, Hirakata, Osaka, Japan.

Abstract

The Cramér's V is popular as an association coefficient in goodness-of-fit tests for contingency tables and its maximum value is known to be 1, but it is not true. We propose a modified Cramér's V.

keyword: Cramér's V

1 Cramér's V

The Cramér's V is popular as an association coefficient in goodness-of-fit tests for contingency tables. For a contingency table

Table 1: Contingency table

	B_1	$\cdots$	B_j	$\cdots$	B_c	sum
A_1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1c}	$x_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ic}	$x_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	x_{r1}	$\cdots$	x_{rj}	$\cdots$	x_{rc}	$x_{r\cdot}$
sum	$x_{\cdot 1}$	$\cdots$	$x_{\cdot j}$	$\cdots$	$x_{\cdot c}$	n

the definition of the Cramér's V is

$$\mathrm{V} \;=\; \sqrt{\frac{\chi^2}{n\,\min(c-1, r-1)}} \tag{1.1}$$

where c is the number of columns, r is the number of rows, and χ^2 is the chi-square statistic of the contingency table as follows:

$$\chi^2 \;=\; \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-e_{ij})^2}{e_{ij}}, \tag{1.2}$$

where e_{ij} is the expectation with respect to the observation x_{ij}. As the probability version for (1.2), [2] wrote, in page 282, that

On the other hand, by means of the inequalities $p_{ij} \le p_{i\cdot}$ and $p_{ij} \le p_{\cdot j}$ it follows from the last expression that $\varphi^2 \le q - 1$, where $q = \min(r, c)$ denotes the smaller of the numbers r and c, or their common value if both are equal.

Note that symbols, etc. above are adapted to the format of this paper and the mean square contingency is

$$\varphi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}, \tag{1.3}$$

and $\varphi^2 = \chi^2 / n$ because of $x_{ij} = n\, p_{ij}$ and $e_{ij} = n\, p_{i\cdot} p_{\cdot j}$.

PROPOSITION 1.1 ([**2**])

$$0 \ \le \ \frac{\varphi^2}{q-1} \ = \ \frac{\varphi^2}{\min(r,c) - 1} \ \le \ 1.$$

The Cramér's V has since been defined as (1.1) whose maximum value has been used as 1. For example, see [1], [3], and [4].

2 The maximum value of V

Since Cramér introduced the contingency coefficient V, its maximum value has been recognized as $n(\min(r, c) - 1)$, as in Proposition 1.1. However, we recognize that this is a mistake and that the correct value is $n(rc - 1)$.

THEOREM 2.1 *For the chi^2 statistic (1.2) of the contingency table, its maximum value is as follows:*

$$\max \chi^2 \ = \ n(rc - 1).$$

LEMMA 2.1 *For the mean square contingency (1.3) of the contingency table, its maximum value is as follows:*

$$\max \varphi^2 \ = \ rc - 1.$$

We first prove the Lemma. The proof of the theorem can be derived naturally from the result.

Table 2: simulation results for Cramér's V and a modified Cramér's V (the number of data is 200 and the number of simulations is 1000.)

	2×2 contingency table		3×3 contingency table	
	V	modified V	V	modified V
Min.	0.0837	0.0483	0.4774	0.2387
1st Qu.	0.7096	0.4097	1.038	0.519
Median	0.9358	0.5403	1.2656	0.6328
Mean	0.9815	0.5667	1.2874	0.6437
3rd Qu.	1.2369	0.7141	1.5263	0.7632
Max.	1.7321	1	2	1

(Proof of Lemma 2.1) By using the relations $p_{ij} \leq p_{i\cdot}$ and $p_{ij} \leq p_{\cdot j}$ that [2] showed, it holds that

$$\begin{aligned}
\varphi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 2\sum_{i=1}^{r}\sum_{j=1}^{c}p_{ij} + \sum_{i=1}^{r}\sum_{j=1}^{c}p_{i\cdot}p_{\cdot j} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1 \\
&\leq \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{p_{i\cdot}p_{\cdot j}}{p_{i\cdot}p_{\cdot j}} - 1 \;=\; \left(\sum_{i=1}^{r}\sum_{j=1}^{c}1\right) - 1 \;=\; rc - 1.
\end{aligned}$$

$\square$

Thus we propose a modified Cramér's V as follows:

$$\text{modified V} = \sqrt{\frac{\chi^2}{n\,(c\,r-1)}} \tag{2.4}$$

3 Simulation

For two contingency tables, 2×2 and 3×3, for the total number $n = 200$, we randomly generate the number of cells. We assume that the probabilities of the cells are all equal and that the number of simulations is 1000.

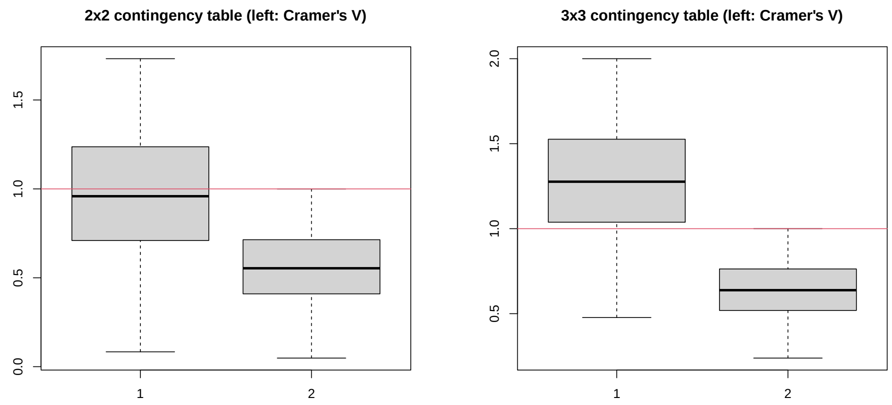

Figure 1: simulation results for Cramér's V and a modified Cramér's V (the number of data is 200 and the number of simulations is 1000.)

4 Conclusion

The proof and simulation results regarding the maximum clearly show that we need to modify the Cramér's V. The previous contingency coefficients must be modified by using a modified Cramér's V with the maximum value 1.

Acknowledgment. This paper was partially supported by Grant-in-Aid for Scientific Research (C) (general) 22K11946 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Haldun Akoglu, User's guide to correlation coefficients, *Turkish Journal of Emergency Medicine*, **18**, 91–93, 2018.

[2] Harald Cramér, *Mathematical methods of statistics*, Princeton University Press, 1946.

[3] Charles C. Okeke, Alternative methods of solving biasedness in Chi-square contingency table, *Academic Journal of Applied Mathematical Sciences*, **5**(1), 1–6, 2019.

[4] Wataru Urasaki, Tomoyuki Nakagawa, Tomotaka Momozaki, and Sadao Tomizawa, Generalized Cramér's coefficient via f-divergence for contin-

gency tables, *Advances in Data Analysis and Classification*, **18**, 893–910, 2024.