A bias-correction for Cramér's V and Tschuprow's T

Wicher Bergsma
London School of Economics and Political Science

Abstract

Cramér's V and Tschuprow's T are closely related nominal variable association measures, which are usually estimated by their empirical values. Although these estimators are consistent, they can have large bias for finite samples, making interpretation difficult. We propose a new and simple bias correction and show via simulations that, for larger than 2×2 tables, the newly obtained estimators outperform the classical (empirical) ones. For 2×2 tables performance is comparable. The larger the table and the smaller the sample size, the greater the superiority of the new estimators.

1 Introduction

Cramér's V is a popular[1] association measure for nominal random variables (Cramér, 1946; Kendall & Stuart, 1973: Chapter 33; Bishop, Fienberg, & Holland, 1975: Chapter 11; Goodman & Kruskal, 1979; Liebetrau, 1983). Closely related and older, but less well-known, is Tschuprow's T (Tschuprow, 1925, 1939), which has some possible theoretical advantages. The usual estimators of these coefficients are simple functions of the Pearson chi-square statistic. Unfortunately, the bias can be very large, especially for small samples, which makes interpretation difficult. In the remainder of this section we define the coefficients. A bias correction is proposed in Section 2, and in Section 3 we show via simulations that the new estimators are superior to the classical ones.

Consider a probability distribution on an $r \times c$ contingency table with the probability in cell (i,j) denoted π_{ij} (i.e., $\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij} = 1$). The mean square contingency (also known as inertia in the correspondence analysis literature) is

$$\phi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}},$$

where a '+' in a subscript denotes summation over that subscript. Two well-known measures

[1] A search for "Cramer's V" receives around 5,000 Google Scholar hits and 50,000 Google hits at the time of writing.

of nominal association based on ϕ^2 are Cramér's V (Cramér, 1946),

$$V = \sqrt{\frac{\phi^2}{\min(r-1, c-1)}},$$

and Tschuprow's T (Tschuprow, 1925, 1939),

$$T = \sqrt{\frac{\phi^2}{\sqrt{(r-1)(c-1)}}}.$$

Both coefficients range from zero to one, with equality to zero if and only if there is independence in the table, i.e., if and only if $\pi_{ij} = \pi_{i+}\pi_{+j}$. Furthermore, $T = 1$ if and only if there is perfect association in the table, i.e., if and only if exactly one cell in each row and each column has nonzero probability. Thus, T can only equal 1 for square tables. On the other hand, V can equal 1 for any rectangular table.

Now consider a multinomial sample of size n on the $r \times c$ contingency table. The proportion of the sample which is in cell (i, j) is denoted p_{ij}. The empirical value $\hat{\phi}^2$ of ϕ^2 is

$$\hat{\phi}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}},$$

and the empirical values of V and T are

$$\hat{V} = \sqrt{\frac{\hat{\phi}^2}{\min(r-1, c-1)}}$$

and

$$\hat{T} = \sqrt{\frac{\hat{\phi}^2}{\sqrt{(r-1)(c-1)}}}.$$

$\hat{V}$ and $\hat{T}$ are the usual estimators, with $\hat{V}$ available in standard statistical packages such as SPSS or SAS. However, it is well known that these are biased estimators of their population counterparts and unfortunately, this bias may be rather large, even for large samples. Consider for example a probability distribution with uniform marginals and satisfying independence on a 5×5 contingency table. In Table 1, the size of simulated biases can be seen for sample sizes $n = 10$, $n = 100$, $n = 1000$ and $n = 10,000$. Even in the latter case, it is still a not entirely negligible 0.02. The standard errors are more than five times smaller than the bias, i.e., most of the root mean squared error (RMSE) is due to the bias. Thus, unless the sample size is large the empirical values $\hat{V}$ and $\hat{T}$ are hard to interpret as measures of strength of association, potentially overestimating the true value by a large amount. Furthermore, since the bias depends on the size of the table and the sample size, the estimators cannot easily be used to compare association in tables of different sizes or of different sample sizes.

In the last row of Table 1 the RMSE of our bias corrected estimator $\tilde{V}$, presented in the next section, is given. It can be seen that the improvement is well above a factor two.

Sample size n	10	100	1000	10,000
Bias of $\hat{V}$	0.609	0.196	0.060	0.018
Standard error of $\hat{V}$	0.089	0.034	0.011	0.003
RMSE of $\hat{V}$	0.616	0.199	0.061	0.019
RMSE of $\tilde{V}$	0.261	0.074	0.023	0.007

Table 1: Simulation results for $\hat{V}$ and $\tilde{V}$ for a 5×5 table with uniform marginals and satisfying independence.

2 Bias correction

The bias of $\hat{\phi}^2$ at independence was conjectured by Tschuprow (1925) and proven by Bartlett (1937) to be

$$E\hat{\phi}^2 = \frac{1}{n-1}(r-1)(c-1). \tag{1}$$

(See also the English translation, Tschuprow, 1939, page 112.) Assuming his then conjecture to be true, Tschuprow suggested the following bias-corrected version of $\hat{\phi}^2$:

$$\tilde{\phi}^2 = \hat{\phi}^2 - \frac{1}{n-1}(r-1)(c-1).$$

Now clearly $\tilde{\phi}^2$ may be negative (or else it could not be unbiased when $\phi^2 = 0$), and since negative values do not make sense, we will instead use the nonnegative estimator

$$\tilde{\phi}^2_+ = \max(0, \tilde{\phi}^2).$$

Note that, since it is nonnegative and not always zero, $\tilde{\phi}^2_+$ is biased at independence.

To obtain bias-corrected estimators of Cramér's V and Tschuprow's T we could just replace the occurrence of $\hat{\phi}^2$ by $\tilde{\phi}^2_+$. However, this would yield a nonzero RMSE when there is perfect dependence in the population. Instead, with

$$\tilde{r} = r - \frac{1}{n-1}(r-1)^2 \qquad \text{and} \qquad \tilde{c} = c - \frac{1}{n-1}(c-1)^2$$

we propose to use

$$\tilde{V} = \sqrt{\frac{\tilde{\phi}^2_+}{\min(\tilde{r}-1, \tilde{c}-1)}}$$

and

$$\tilde{T} = \sqrt{\frac{\tilde{\phi}^2_+}{\sqrt{(\tilde{r}-1)(\tilde{c}-1)}}}.$$

It is now straightforward to verify that, with probability 1, $T = 1$ implies $\tilde{T} = 1$ and $V = 1$ implies $\tilde{V} = 1$.

Due to the bias correction, RMSE can be expected to be small if $\phi^2 = 0$; simulation results given in Table 1 show that in the above example of a 5×5 contingency table, the RMSE for $\tilde{V}$ resp. $\tilde{T}$ is less than half that of $\hat{V}$ resp. $\tilde{T}$ in each of the presented cases. In the next section, simulations show that also when there is association in present in a table (i.e., $\phi^2 > 0$), $\tilde{V}$ and $\tilde{T}$ are superior to $\hat{V}$ resp. $\hat{T}$.

3 Simulations

In this section we compare the bias and RMSE of $\hat{V}^2$ and $\tilde{V}^2$ through simulations. (The reason we do not compare bias and RMSEs of $\hat{V}$ and $\tilde{V}$ is that we found these to have very differently shaped distributions, making comparison of bias and RMSE inappropriate, see Cox & Hinkley, 1994, Section 8.5. For the comparison of $\hat{V}^2$ and $\tilde{V}^2$, the RMSE should be a reasonable indicator of performance.)

The simulations are for square tables so all results are identical for Cramér's V and Tschuprow's T. Hence, to simplify presentation, we omit reference to Tschuprow's T.

By design, the bias correction should work well if independence holds in the table. Hence, we first investigate whether the bias correction also works if there is dependence. For this we will do simulations for a "diagonal association" model for $r \times r$ tables, for which all diagonal cells have the same probability and all off-diagonal cells have the same probability. We parameterize the model using a parameter $\theta \in [-\frac{1}{r-1}, 1]$, with the diagonal cells having probability $\frac{1}{r^2} + \frac{\theta(r-1)}{r^2}$ and the off-diagonal cells having probability $\frac{1}{r^2} - \frac{\theta}{r^2}$. It is immediately seen that independence holds if $\theta = 0$. With this parameterization, it can be shown that

$$V = |\theta|.$$

For a 3×3 table, the model is given as

$\frac{1}{9} + \frac{2\theta}{9}$	$\frac{1}{9} - \frac{\theta}{9}$	$\frac{1}{9} - \frac{\theta}{9}$
$\frac{1}{9} - \frac{\theta}{9}$	$\frac{1}{9} + \frac{2\theta}{9}$	$\frac{1}{9} - \frac{\theta}{9}$
$\frac{1}{9} - \frac{\theta}{9}$	$\frac{1}{9} - \frac{\theta}{9}$	$\frac{1}{9} + \frac{2\theta}{9}$

.

Figures 1, 2, and 3 give simulation results for $\theta \in [-\frac{1}{r}, 1]$ for $r \times r$ tables with $r = 3$, $r = 5$, and $r = 7$, respectively, and sample sizes $n = 20$, $n = 100$, and $n = 1000$. It can be seen that the bias is reduced significantly for all values of θ.

In Figures 4, 5, and 6 the RMSE of $\hat{V}^2$ and $\tilde{V}^2$ for the same diagonal association model is plotted. It can be seen that the new estimator gives the most improvement for moderate levels of association and small sample sizes. The improvement also increases somewhat with table size.

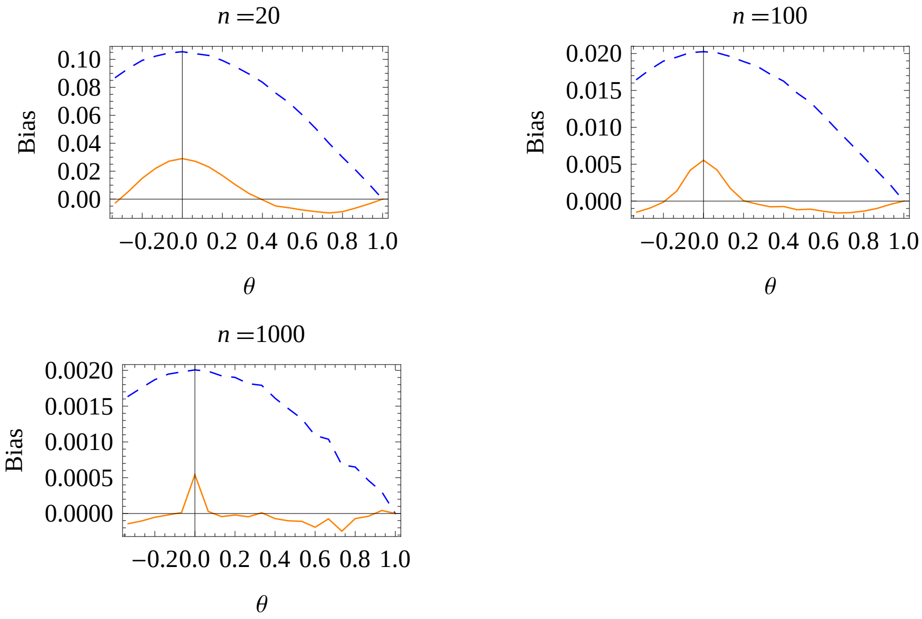

Figure 1: Bias of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 3×3 table under the diagonal association model.

It may be remarked that in Figures 4, 5, and 6 the 'bumps' in each case becomes more pronounced as the sample size increases, in particular, at $\theta = 0$ (independence), the RMSE decreases at a rate of approximately $1/n$, and for largish values of θ at a rate of approximately $1/\sqrt{n}$. To see why this is so, recall that $n \times \min(r-1, c-1)\hat{V}^2$ is the Pearson chi-square statistic which has an asymptotic chi-square distribution under independence, so $\hat{V}^2 - \theta = O_p(1/n)$ if $\theta = 0$, while by asymptotic normality $\hat{V}^2 - \theta = O_p(1/\sqrt{n})$ if $0 < \theta < 1$.

In Figure 7, we give the ratio of the average RMSEs of $\hat{V}^2$ and $\tilde{V}^2$ for sample sizes $n = 20$ and $n = 100$ for up to 7×7 tables. The average is computed over 10,000 probability distributions sampled from a uniform distribution over the $(rc-1)$-dimensional probability simplex (note that this is a Dirichlet distribution with concentration vector $(1, 1, \ldots, 1)'$). On average, for a 7×7 table, the RMSE of $\hat{V}^2$ is about three times larger than the RMSE of $\tilde{V}^2$. We see that, on average, the bias-corrected estimator is always better than the classical one. The smaller the sample, the greater the relative difference, and also the larger the table, the greater the relative difference.

We also investigated, but did not present here, the case of a 2×2 table, and found there to be little difference in RMSE of the two estimators. Note that, for a 2×2 table, Cramér's V equals the square of the correlation coefficient.

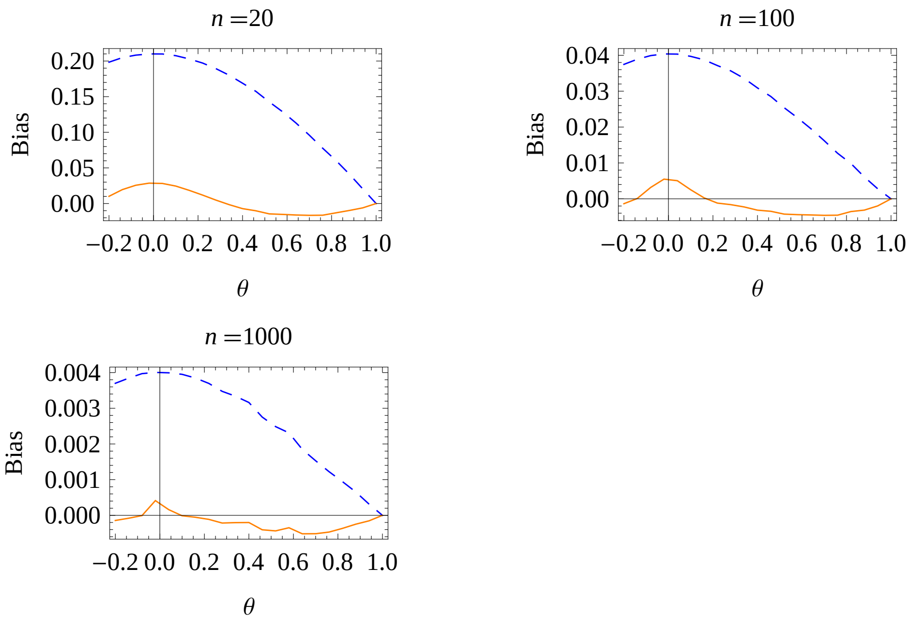

Figure 2: Bias of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 5×5 table under the diagonal association model.

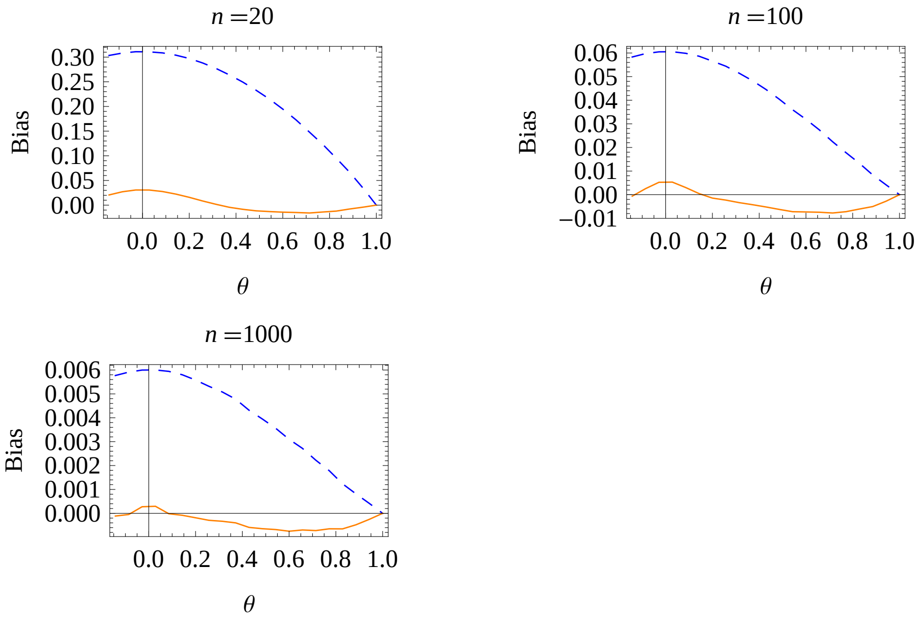

Figure 3: Bias of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 7×7 table under the diagonal association model.

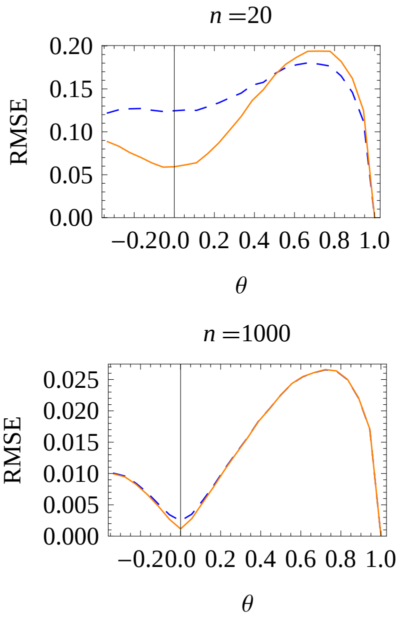

Figure 4: RMSE of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 3×3 table under the diagonal association model.

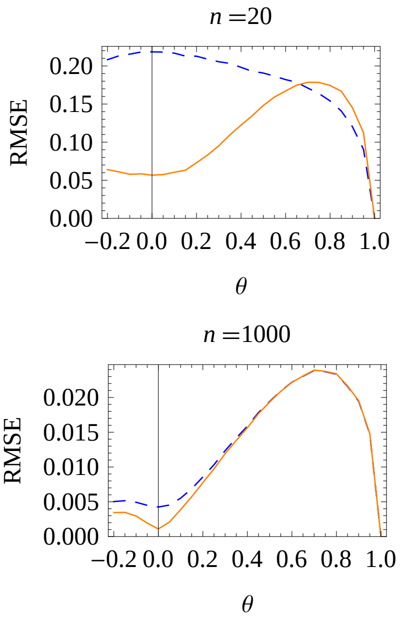

Figure 5: RMSE of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 5×5 table under the diagonal association model.

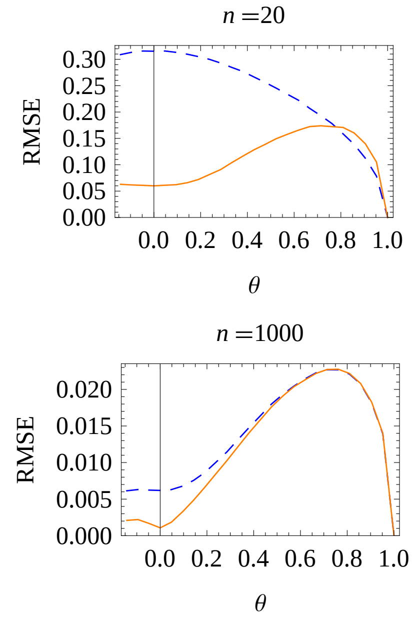

Figure 6: RMSE of empirical estimator $\hat{V}^2$ (blue dashed line) and bias-corrected estimator $\tilde{V}^2$ (orange solid line) of Cramér's V for a 7×7 table under the diagonal association model.

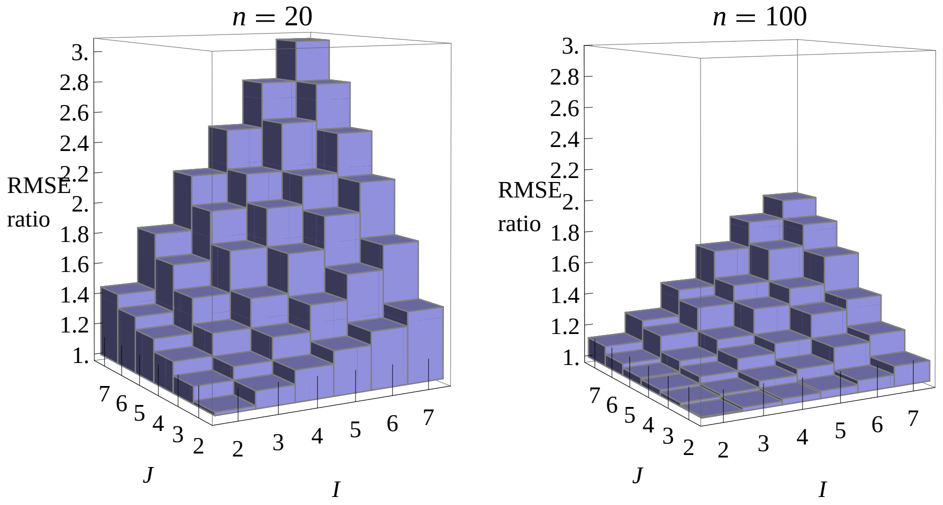

Figure 7: Ratio of average RMSEs of empirical value versus bias-corrected estimator of Cramér's V for $I \times J$ tables for $n = 20$ and $n = 100$. It is seen that the larger the table and the smaller the sample size, the better the relative performance of the bias-corrected estimator.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *160*(901), 268-282.

Bishop, Y. V. V., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

Cox, D., & Hinkley, D. (1994). *Theoretical statistics.* Chapman & Hall.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton Press, NJ.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications.* New York: Springer-Verlag.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics. Vol. 2* (Third ed.). Hafner Publishing Co., New York. (Inference and relationship)

Liebetrau, A. (1983). *Measures of association.* Sage Publications, Inc.

Tschuprow, A. (1925). *Grundbegriffe und grundprobleme der korrelationstheorie.* Leipzig: B.G. Teubner.

Tschuprow, A. (1939). *Principles of the mathematical theory of correlation (translation of Tschuprow, 1925, by M. Kantorowitsch).* W. Hodge & Co.