Analysis of statistical association of two dichotomous variables with a 2×2 contingency table

Martin Förster

September 28, 2026

Statistical association

This note concerns the analysis of the statistical association between two dichotomous variables - *without any assumption of causality*. Therefore, in the following, "X" and "Y" do *not* refer to independent and dependent variables, respectively, but simply to two variables.

Example data

Table 1 shows the bi-variate distribution of the example data with a 2×2 cross table. Both variables are originally treated as metric and have been dichotomized here. More on this later.

Statistics with data at the individual level

One version of Rho (ϱ) is a fit- and parameter-based effect size that shows the signed contribution to the deviation from the null model. This Rho is a standardized measure of how non-random the signed description of the statistical

Table 1: Example Data

		X		
		1	0	$\sum$
Y	1	^{a} 787	^{b} 1665	2452
	0	^{c} 382	^{d} 541	923
	$\sum$	1169	2206	3375

association is. It adjusts for sample size: since the standard error (SE) gets smaller the larger the sample, for the same effect size, you need to explain more variance in larger samples than in smaller ones. In this case,

$$\varrho = -0.08$$

This means that a negative correlation ($\frac{a}{a+c} < \frac{b}{b+d}$ and $\frac{a}{a+b} < \frac{c}{c+d}$) is identified, which, however, is negligibly weak.

Another version of Rho is a variance-based effect size along with a chi-squared test statistic. This shows the signed contribution (ϱ_{Var}) and how statistically significant (χ^2) the deviation from the null model is. The residuals, whose variance is calculated, are obtained using

$$e_i\,|a = \frac{P(a)}{4P(Y)P(X)}$$

$$e_i\,|b = -\frac{P(b)}{4P(Y)(1-P(X))}$$

$$e_i\,|c = -\frac{P(c)}{4(1-P(Y))P(X)}$$

$$e_i\,|d = \frac{P(d)}{4(1-P(Y))(1-P(X))}$$

The standardized ϱ_{Var} is a measure of how much larger the empirical variance (of the residuals) is than the expected variance under the assumption of independence. The sign of ϱ_{Var} reflects an over- or under-population of the cells (compared to independence): if ϱ_{Var} is positive, it indicates an over-population of cells a and d; a negative sign, on the other hand, indicates an over-population of cells b and c. Here the result is

$$\varrho_{Var} = -0.28 \quad (p < 0.001)$$

and

$$\chi^2 = 3671.49 \quad (df = 3374;\, p < 0.001)$$

The result here is that both the deviation from the null model and the negative relationship is significant.

The ϱ_{Var} gets its sign from the mean of the residuals. It is also this mean whose deviation from zero is tested for statistical significance. In this example, $\overline{e} = -0.07$. Thus, $\overline{e}$ is numerically and conceptually closer to ϱ. A normalization of $\overline{e}$ can be achieved by multiplying it by 2. That is, $2\overline{e}$ and ϱ describe the strength of the relationship, ϱ_{Var} describes the extend of deviation from the null model.

Statistics with aggregated data

With aggregated data, an effect size for the association can easily be calculated from a 2×2 contingency table: the odds ratio. The odds ratio is $OR = \frac{a/c}{b/d}$, which would be $OR = 1$ if the two variables were independent. This measure is not affected by the sample size or the marginal distribution. For the exemplary 2×2 contingency table,

$$OR = 0.67 \quad 95\%CI = [0.59\,;\,0.80]$$

This result indicates a negative, statistically significant association of the variables.

Finally, the contingency chi-square ($df = 1$) is a prominent measure of deviation from the null model, i.e., of the deviation of the empirical distribution (cell frequencies) from the distribution expected under independence. The contingency chi-square is influenced by sample size and marginal distribution. For the example data, we obtain

$$\chi^2 = 25.57 \quad (p < 0.001)$$

The test statistic for the contingency indicates a significant deviation from independence and thus a statistical association between the two variables.

Notes on dichotomization

Both variables were originally metrically discretely scaled, Y with an (integer) range of values from 1 to 7, X from 1 to 10. Although there are generally some concerns regarding dichotomization, under certain conditions - some of which are met here - the advantages outweigh these concerns. Both variables have a limited range of values, which could lead to the parameters of a linear estimation model (for which the original, non-dichotomous variables would be used) being estimated with bias. However, a linear estimation model rarely provides added value, at least in the social sciences, because and insofar as the hypotheses being tested rarely formulate linear relationships, but generally monotonic associations. Categorization would thus only lead to the loss of information that was not sought or can only be interpreted on an ad-hoc basis. If such information was already relevant when coming up with the hypothesis, you wouldn't have to test the null hypothesis for sure, and might go straight for the alternative hypothesis.

With this reasoning and a possibly justified categorization, an ordinal estimation model would, in principle, still be appropriate. However, if - as in this example - some categories are very sparsely populated, even ordinal models carry a risk of unstable estimates. For Y, this concerns the categories/values 1 (1.6%), 2 (2.6%), and 3 (7%). Combining just these three categories would, however, be arbitrary in terms of meaning and would have only a statistical rationale. In contrast, dividing or grouping the value ranges 1..4 and 5..7 makes

Table 2: Statistics with simulated data (part 1/2)

Simulation setup	Sample size	
	$n = 100$	$n = 10000$
I (sym: random noise only)	28 (28%) | 18 (18%) 26 (26%) | 28 (28%)	2486 (25%) | 2497 (25%) 2541 (25%) | 2476 (25%)
	$\varrho = 0.12$	$\varrho = -0.01$
	$\varrho_{Var} = 0.24 \quad (\overline{e}_\varrho = 0.06;\ p = 0.006)$ $\chi^2 = 105.08 \quad (df = 99;\ p > 0.1)$ $OR = 1.68 \quad 95\%CI = [0.75\,,\ 3.94]$ $\chi^2_C = 1.62 \quad (df = 1;\ p > 0.1)$	$\varrho_{Var} = -0.02 \quad (\overline{e}_\varrho = -0.004;\ p = 0.07)$ $\chi^2 = 10001.3 \quad (df = 9999;\ p > 0.1)$ $OR = 0.97 \quad 95\%CI = [0.87\,,\ 1.05]$ $\chi^2_C = 0.57 \quad (df = 1;\ p > 0.1)$
II (sym: extreme positive effect)	41 (41%) | 5 (5%) 2 (2%) | 52 (52%)	4473 (45%) | 510 (5%) 490 (5%) | 4527 (45%)
	$\varrho = 0.04$	$\varrho = 0.67$
	$\varrho_{Var} = 0.95 \quad (\overline{e}_\varrho = 0.43;\ p < 0.001)$ $\chi^2 = 1074.27 \quad (df = 99;\ p < 0.001)$ $OR = 213.2 \quad 95\%CI = [44.2\,,\ +\infty]$ $\chi^2_C = 73.96 \quad (df = 1;\ p < 0.001)$	$\varrho_{Var} = 0.94 \quad (\overline{e}_\varrho = 0.4;\ p < 0.001)$ $\chi^2 = 81085 \quad (df = 9999;\ p < 0.001)$ $OR = 81.03 \quad 95\%CI = [69.24\,,\ 95.21]$ $\chi^2_C = 6400.02 \quad (df = 1;\ p < 0.001)$

sense. For X, the distribution of the categories 1 (3.7%), 2 (5%), 8 (4.5%), 9 (1.3%), and 10 (1.1%) is critical. Here too, merging the two lower and three upper extreme categories would be arbitrary, or perhaps appropriate only for this particular data set, and thus hardly comparable.

Examples with simulated data

Table 2 and table 3 reports the simulated contingency tables and the statistical results. The configuration of the contingency tables is | a | b | / | c | d |. Since both ϱ and the odds ratio CI where calculated bootstrap based, the combination of a very extreme distribution by a low n (setup II, $n = 100$) could result in biased

Table 3: Statistics with simulated data (part 2/2)

Simulation setup	Sample size	
	$n = 100$	$n = 10000$
III (sym: moderate negative effect)	20 (20%) | 26 (26%) 37 (37%) | 17 (17%)	1482 (15%) | 3501 (35%) 3492 (35%) | 1525 (15%)
	$\varrho = -0.24$	$\varrho = -0.31$
	$\varrho_{Var} = -0.47 \quad (\overline{e}_\varrho = -0.13;\ p < 0.001)$ $\chi^2 = 126.58 \quad (df = 99;\ p = 0.025)$ $OR = 0.35 \quad 95\%CI = [0.15\,,\ 0.82]$ $\chi^2_C = 6.35 \quad (df = 1;\ p = 0.012)$	$\varrho_{Var} = -0.66 \quad (\overline{e}_\varrho = -0.2;\ p < 0.001)$ $\chi^2 = 17553.99 \quad (df = 9999;\ p < 0.001)$ $OR = 0.18 \quad 95\%CI = [0.17\,,\ 0.2]$ $\chi^2_C = 1589.02 \quad (df = 1;\ p < 0.001)$
IV (asym: negative, no noise)	25 (25%) | 47 (47%) 21 (21%) | 7 (7%)	2500 (25%) | 4700 (47%) 2100 (21%) | 700 (7%)
	$\varrho = -0.14$	$\varrho = -0.33$
	$\varrho_{Var} = -0.58 \quad (\overline{e}_\varrho = -0.17;\ p < 0.001)$ $\chi^2 = 149.75 \quad (df = 99;\ p < 0.001)$ $OR = 0.18 \quad 95\%CI = [0.06\,,\ 0.47]$ $\chi^2_C = 13.17 \quad (df = 1;\ p < 0.001)$	$\varrho_{Var} = -0.58 \quad (\overline{e}_\varrho = -0.17;\ p < 0.001)$ $\chi^2 = 15124.95 \quad (df = 9999;\ p < 0.001)$ $OR = 0.18 \quad 95\%CI = [0.16\,,\ 0.19]$ $\chi^2_C = 1316.65 \quad (df = 1;\ p < 0.001)$

estimates respectively.

Supplement

https://foerster-m.de/Nexus/databunker/2x2_contingency/dict_romod.html