mathematics

MDPI

Article

Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data that Use the Chi-Squared Statistic

Mattan S. Ben-Shachar [1,*], Indrajeet Patil [2], Rémi Thériault [3], Brenton M. Wiernik [4], and Daniel Lüdecke [5]

[1] Independent researcher; mattansb@msbstats.info
[2] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany; patilindrajeet.science@gmail.com
[3] Department of Psychology, Université du Québec à Montréal, Montréal, Québec, Canada; theriault.remi@courrier.uqam.ca
[4] Independent researcher; brenton@wiernik.org
[5] Institute of Medical Sociology, University Medical Center Hamburg-Eppendorf, Germany; d.luedecke@uke.de

* Correspondence: mattansb@msbstats.info

Abstract: In both theoretical and applied research, it is often of interest to assess the strength of an observed association. Existing guidelines also frequently recommend going beyond null-hypothesis significance testing and to report effect sizes and their confidence intervals. As such, measures of effect sizes are increasingly reported, valued, and understood. Beyond their value in shaping the interpretation of the results from a given study, reporting effect sizes is critical for meta-analyses, which rely on their aggregation. We here review the most common effect sizes for analyses of categorical variables that use the χ^2 (chi-square) statistic, and introduce a new effect size—פ (Fei, pronounced "fay"). We demonstrate the implementation of these measures and their confidence intervals via the *effectsize* package in the R programming language.

Keywords: Effect Sizes; Chi-Squared Test; Phi; Cramer's V, Fei

MSC: 62-01; 62-04; 62-08; 62H17; 62P99

1. Introduction

Over the last two decades, there has been growing concerns about the so-called replication crisis in psychology and other fields [1,2]. As a result, the scientific community has paid increasing attention to the issue of replicability in science, as well as to good research and statistical practices.

In this context, many have highlighted the limitations of null-hypothesis significance testing and called for more modern approaches to statistics. One such recommendation coming for example from the "New Statistics" initiative is to report effect sizes and their corresponding confidence intervals, and to increasingly rely on meta-analyses to increase confidence in those estimations [3]. These recommendations are meant to complement (or even replace, according to some) null-hypothesis significance testing and would help transition toward a "cumulative quantitative discipline".

These so-called "New Statistics" are synergistic because effect sizes are not only useful for interpreting study results in themselves, but also because they are necessary for meta-analyses, which aggregate effect sizes and their confidence intervals to create a summary effect size of its own [4,5].

Unfortunately, popular software applications do not always offer the necessary implementations of the specialized effect sizes necessary for a given research design and

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

their confidence intervals. In this paper, we want to focus on effect sizes for categorical data that are probably less well known than popular effect sizes like Cohen's *d* or Pearson's *r* [6,7]. For categorical data, *d* and *r* are inappropriate measures of an effect size. Cohen's *d* refers to the standardized difference between the means of two populations, while Pearson's correlation coefficient *r* measures linear correlations. Hence, both measures refer to continuous data, not categorical.

To compare categorical data, for instance, where associations can be presented as contingency tables, several effect size metrics are available. Common effect sizes for 2-by-2 tables are odds ratios (OR), risk ratios (RR) or the *phi* (ϕ) coefficient. While *phi* can be interpreted similarly to a correlation coefficient, OR and RR are harder to interpret as they are not bounded between zero and one. Furthermore, RR are not symmetrical. The effect size can change when columns and rows are exchanged. For tables with larger dimensions than 2-by-2, other effect sizes (like Cramér's *w*) are available that share the property of *phi* of being able to be interpreted like a correlation coefficient and which are discussed later.

The *observed* distribution of categorical data—usually measured as multinomial variables—can also be compared to an *expected* distribution. Again, effect sizes to measure the strength of such associations show some limitations regarding the ease of interpretation. What is missing here is an effect size which metric is comparable to those for contingency tables.

The aim of this paper is to review the most commonly used effect sizes for analyses of categorical variables that use the χ^2 (chi-square) test statistic, and introduce a new effect size, פ (Fei, pronounced "fay"), which closes the gap of a missing effect size measure in a correlation like metric that is appropriate for categorical data.

Importantly, we offer researchers an applied walkthrough on how to use these effect sizes in practice thanks to the *effectsize* package in the R programming language, which implements these measures and their confidence intervals [8,9]. The presented *effectsize* package closes another gap related to before mentioned effect sizes, because the uncertainty of such measures—expressed by their confidence intervals—is often not included in the output of statistical software. We cover in turn tests of independence (φ/phi, Cramér's *V*) and tests of goodness of fit (Cohen's *w*, Tschuprow's *T* and a new proposed effect size, פ/Fei).

2. Tests of Independence

The χ^2 test of independence between two categorical variables examines if the frequency distribution of one of the variables is dependent on the other. That is, are the two variables correlated such that, for example, members of group 1 on variable X are more likely to be members of group A on variable Y, rather than evenly spread across Y variable groups A and B. Formally, the test examines how likely the observed conditional frequencies (cell frequencies) are under the null hypotheses of independence. This is done by examining the degree the observed cell frequencies deviate from the frequencies that would be expected if the variables were indeed independent. The test statistic for these tests is the χ^2, which is computed as:

$$\chi^2 = \sum_{i=1}^{l \times k} \frac{(O_i - E_i)^2}{E_i}$$

Where O_i are the *observed* frequencies and E_i are the frequencies *expected* under independence, and l and k are the number of rows and columns of the contingency table.

Instead of the deviations between the observed and expected frequencies, we can write χ^2 in terms of observed and expected cell *probabilities* and the total sample size N (since $p = k/N$):

$$\chi^2 = N \times \sum_{i=1}^{l \times k} \frac{\left(p_{O_i} - p_{E_i}\right)^2}{p_{E_i}}$$

Where p_{O_i} are the *observed* cell probabilities and p_{E_i} are the probabilities *expected* under independence.

Table 1 gives a short example in R to demonstrate whether the probability of survival of the sinking of the Titanic is dependent on the sex of the passenger. The null hypothesis tested here is that the probability of survival is independent of the passenger's sex.

Table 1. $\chi 2$ test of survival of Titanic passengers by sex, Titanic dataset from R

Sex	Survived	Died
Male	367	1364
Female	344	126

$\chi^2 = 454.5$, $df = 1$, $p < 0.001$***

These results can be reproduced with the following R code:

```
(Titanic_xtab <- as.table(apply(Titanic, c(2, 4), sum)))
chisq.test(Titanic_xtab)
```

The performed χ^2-test is statistically significant, thus we can reject the hypothesis of independence. However, the output includes no effect size. We cannot draw conclusions of the strength of the association between sex and survival.

2.1 Phi

For a 2-by-2 contingency table analysis, as the one used above, the ϕ (*phi*) coefficient is a correlation-like measure of effect size indicating the strength of association between the two binary variables. One possibility to compute this effect size is to re-code the binary variables as dummy ("0" and "1") variables, and computing the Pearson correlation between them [10]:

$$\phi = |r_{AB}|$$

Another way to compute ϕ is by using the χ^2 statistic:

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\sum_{i=1}^{l \times k} \frac{\left(p_{O_i} - p_{E_i}\right)^2}{p_{E_i}}}$$

This value ranges between zero (no association) and one (complete dependence), and its values can be interpreted the same as Person's correlation coefficient. Table 2 shows the correlation coefficient and the effect size ϕ for the data shown in table 1.

Table 2. Correlation and effect size ϕ (*phi*) for the survival of Titanic passengers by sex, Titanic dataset from R

Variable 1	Variable 2	*r* (95% CI)	ϕ (95% CI)
Sex (male/female)	Survival (survived/died)	-.46 (-.49, -.42)	.46 (.42, 1.00)

ϕ can be estimated with the *effectsize* package in R using the following function:

```
effectsize::phi(Titanic_xtab, adjust = FALSE)
```

Note that ϕ cannot be negative, so will take the *absolute* value of Pearson's correlation coefficient. Also note that the *effectsize* package gives a *one-sided* confidence interval by default, to match the positive direction of the associated χ^2 test at $\alpha = .05$ (that the association is *larger* than zero at a 95% confidence level).

2.2 Cramér's V *(and Tschuprow's* T*)*

When the contingency table is larger than 2-by-2, using $\sqrt{\chi^2/N}$ can produce values larger than one, and so loses its interpretability as a correlation-like effect size. Cramér showed that while for 2-by-2 the maximal possible value of χ^2 is N, for larger tables the maximal possible value for χ^2 is $N \times (\min(k, l) - 1)$ [11]. Therefore, he suggested the V effect size (also sometimes known as Cramér's phi and denoted as ϕ_c):

$$\text{Cramer's } V = \sqrt{\frac{\chi^2}{N(\min(k, l) - 1)}}$$

V is one when the columns are completely dependent on the rows, or the rows are completely dependent on the columns (and zero when rows and columns are completely independent).

Table 3. Effect size Cramér's *V* for of survival of Titanic passengers by sex, Titanic dataset from R

Class/Position	Survived	Died
1st	203	122
2nd	118	167
3rd	178	528
Crew	212	673

Cramér's *V* = .29, 95% CI = .26, 1.00

These results can be reproduced with the following R code:

```
(Titanic_xtab2 <- as.table(apply(Titanic, c(1, 4), sum)))
effectsize::cramers_v(Titanic_xtab2, adjust = FALSE)
```

Tschuprow devised an alternative value, at

$$\text{Tschuprow's } T = \sqrt{\frac{\chi^2}{N\sqrt{(k-1)(l-1)}}}$$

which is one only when the columns are completely dependent on the rows *and* the rows are completely dependent on the columns, which is only possible when the contingency table is a square [12].

For example, in the following table, each row is dependent on the column value; that is, if we know if the food is a soy, milk, or meat product, we also know whether the food is vegan or not. However, the columns are *not* fully dependent on the rows: knowing the food is vegan tells us the food is soy based, however, knowing it is not vegan does not allow us to classify the food—it can be either a milk product or a meat product.

Accordingly, as can be seen in Table 4, Cramer's V will be one, but Tschuprow's T will not be:

Table 4. Cramér's *V* and Tschuprow's *T* for food classes, example dataset from R

Type	Product			Cramér's V (95% CI)	Tschuprow's T (95% CI)
	Soy	Milk	Meat		
Vegan	47	0	0	1.00 (.81, 1.00)	.84 (.68, 1.00)
Not-Vegan	0	12	12		

These results can be reproduced with the following R code:

```
data("food_class", package = "effectsize")
effectsize::cramers_v(food_class, adjust = FALSE)
effectsize::tschuprows_t(food_class)
```

We can generalize ϕ, V, and T to: $\sqrt{\frac{\chi^2}{\chi^2_{\max}}}$. That is, they express a proportional of the sample-χ^2 to the maximally possible χ^2 given the study design.

These coefficients can also be used for confusion matrices, which are 2-by-2 contingency tables used in assessing machine learning algorithms classification abilities, comparing true outcome classes with the model-predicted outcome class. In fact, a popular metric is the Matthews correlation coefficient (MCC) for binary classifiers, which is often presented in terms of true and false positives and negatives, is nothing more than ϕ [13].

3. Goodness of Fit

These tests compare an observed distribution of a multinomial variable to an expected distribution, using the same χ^2 statistic. Here, in addition, we can compute an effect size as $\sqrt{\frac{\chi^2}{\chi^2_{\max}}}$; all we need to find is $\chi^2_{\max}$.

3.1 Cohen's w

Cohen (Cohen 2013) defined an effect size—w—for the goodness of fit test:

$$\text{Cohen's } w = \sqrt{\sum_{i=1}^{k} \frac{\left(p_{O_i} - p_{E_i}\right)^2}{p_{E_i}}} = \sqrt{\frac{\chi^2}{N}}$$

Thus, $\chi^2_{\max} = N$.

Unfortunately, w has an upper bound of one *only* when the variable is binomial (has two categories) and the expected distribution is uniform ($p = 1 - p = 0.5$). When the distribution is non-uniform or if there are more than two classes, then $\chi^2_{\max} > N$, and so w can be larger than one [14,15]. Examples are shown in Table 5.

Table 5. Effect size Cohen's w for variables with different number of categories and distributions

Observed counts	Expected proportion	Cohen's w (95% CI)
90 / 10	.5 / .5	0.80 (0.61, 1.00)
90 / 10	.35 / .65	1.15 (0.99, 1.36)
5 / 10 / 80 / 5	.25 / .25 / .25 / .25	1.27 (1.10, 1.73)

These results can be reproduced with the following R code:

```
O <- c(90, 10)
effectsize::cohens_w(O, p = c(0.5, 0.5))
effectsize::cohens_w(O, p = c(0.35, 0.65))
```

```
O <- c(5, 10, 80, 5)
effectsize::cohens_w(O, p = c(0.25, 0.25, 0.25, 0.25))
```

Although Cohen suggested that w can also be used for such designs, we believe that this hinders the interpretation of w since it can be arbitrarily large [7].

3.2 Fei

We present here a new effect size, פ (Fei, pronounced "fay"), which normalizes goodness-of-fit χ^2 by the proper $\chi^2_{\max}$ for non-uniform and/or multinomial variables.

The largest deviation from the expected probability distribution would occur when all observations are in the cell with the smallest expected probability. That is:

$$p_O = \begin{cases} 1, & \text{if } p_i = \min(p) \\ 0, & \text{Otherwise} \end{cases}$$

We can find $\frac{(E_i - O_i)^2}{E_i}$ for each of these values:

$$\frac{(p_E - p_O)^2}{p_E} = \begin{cases} \frac{(p_i - 1)^2}{p_i} = \frac{(1 - p_i)^2}{p_i}, & \text{if } p_E = \min(p_E) \\ \frac{(p_i - 0)^2}{p_i} = p_i, & \text{Otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{k} \frac{\left(p_{O_i} - p_{E_i}\right)^2}{p_{E_i}} &= \sum_{i=1}^{k} p_{E_i} - \min(p_E) + \frac{\left(1 - \min(p_E)\right)^2}{\min(p_E)} \\ &= 1 - \min(p_E) + \frac{\left(1 - \min(p_E)\right)^2}{\min(p_E)} \\ &= \frac{1 - \min(p_E)}{\min(p_E)} \\ &= \frac{1}{\min(p_E)} - 1 \end{aligned}$$

And so,

$$\begin{aligned} \chi^2_{\max} &= N \times \sum_{i=1}^{k} \frac{\left(p_{O_i} - p_{E_i}\right)^2}{p_{E_i}} \\ &= N \times \left(\frac{1}{\min(p_E)} - 1\right) \end{aligned}$$

Finally, an effect size can be derived as:

$$\sqrt{\frac{\chi^2}{N \times \left(\frac{1}{\min(p_E)} - 1\right)}}$$

We call this effect size פ (Fei), which represents the voiceless bilabial fricative in the Hebrew language, keeping in line with ϕ (which in modern Greek marks the same sound) and V (which in English marks a voiced bilabial fricative; W being derived from

the letter V in modern Latin alphabet). ₰ will be zero when the observed distribution perfectly matches the one expected (under the null hypothesis), and will be one when the sample contains *only* one class of observations—the one with the smallest expected probability (under the null hypothesis). That is, ₰ is 1 (its maximal value) only when we observe only the least expected class. When there are only two cells with uniform expected probabilities (50%), the expression $N \times \left(\frac{1}{\min(p_E)} - 1\right)$ reduces to N and so ₰ $= w$. Table 6 shows the effect size Fei for the same vectors and distributions as seen for Cohen's w in Table 5. As can be seen, unlike Cohen's w, all effect size values of Fei (and their confidence intervals) are within the range from zero to one.

Table 6. Effect size Fei for variables with different number of categories and distributions

Observed counts	Expected proportion	Fei (95% CI)
90 / 10	.5 / .5	.80 (.64, 1.00)
90 / 10	.35 / .65	.85 (.73, 1.00)
5 / 10 / 80 / 5	.25 / .25 / .25 / .25	.73 (.64, 1.00)

These results can be reproduced with the following R code:

```
O <- c(90, 10)
effectsize::fei(O, p = c(0.5, 0.5))
effectsize::fei(O, p = c(0.35, 0.65))

O <- c(5, 10, 80, 5)
effectsize::fei(O, p = c(0.25, 0.25, 0.25, 0.25))
```

4. Conclusion

Effect sizes are essential to interpret the magnitude of observed effects, they are frequently required in scientific journals, and they are necessary for a cumulative quantitative science relying on meta-analyses. In this paper, we have covered the mathematics and implementation in R of four different effect sizes for analyses of categorical variables that specifically use the χ^2 (chi-square) statistic. Furthermore, with our proposal of the effect size ₰ (Fei), we fill the missing effect size for all cases of a χ^2 test.

Table 7. Effect size for χ^2 tests for differently sized contingency tables

Test	Table Size	Effect size
χ^2 test for independence	2-by-2	ϕ
	Larger than 2-by-2	V or T (Reduces to ϕ when table is 2-by-2)
χ^2 test for goodness-of-fit	2 classes with uniform null distribution	w
	More than 2 classes and/or non-uniform null distribution	₰ (Reduces to w when there are 2 classes with uniform null dist.)

Thus, we now have effect sizes to accompany any sized 1-dimentional or 2-dimentional contingency tables, that represent the sample's χ^2 relative to the maximally possible χ^2, ranging from 0 to 1, that can be easily interpreted on the scale of a correlation coefficient.

Author Contributions: M.S.B. conceptualized and developed the Fei effect size and its implementation in *effectsize*, and drafted the paper; all authors contributed to both the writing of the paper and the conception of the software. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The R code to reproduce the results from the tables in this article can downloaded at https://osf.io/cg64s/ (doi: 10.17605/osf.io/cg64s).

Acknowledgments: The {effectsize} package is part of the collaborative R *easystats* ecosystem. Thus, we thank all members of *easystats*, contributors, and users alike.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Open Science Collaboration Estimating the Reproducibility of Psychological Science. *Science* **2015**, *349*, aac4716, doi:10.1126/science.aac4716.
2. Camerer, C.F.; Dreber, A.; Holzmeister, F.; Ho, T.-H.; Huber, J.; Johannesson, M.; Kirchler, M.; Nave, G.; Nosek, B.A.; Pfeiffer, T.; et al. Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2018**, *2*, 637–644, doi:10.1038/s41562-018-0399-z.
3. Cumming, G. The New Statistics: Why and How. *Psychol. Sci.* **2014**, *25*, 7–29, doi:10.1177/0956797613504966.
4. Wiernik, B.M.; Dahlke, J.A. Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Adv. Methods Pract. Psychol. Sci.* **2020**, *3*, 94–123, doi:10.1177/2515245919885611.
5. DeGeest, D.S.; Schmidt, F.L. The Impact of Research Synthesis Methods on Industrial-Organizational Psychology: The Road from Pessimism to Optimism about Cumulative Knowledge. *Res. Synth. Methods* **2010**, *1*, 185–197, doi:10.1002/jrsm.22.
6. Pearson, K. VII. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242, doi:10.1098/rspl.1895.0041.
7. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; 2nd ed.; Routledge, 1988; ISBN 978-0-203-77158-7.
8. Ben-Shachar, M.; Lüdecke, D.; Makowski, D. Effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J. Open Source Softw.* **2020**, *5*, 2815, doi:10.21105/joss.02815.
9. R Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023;
10. Olvera Astivia, O.L. The Relationship between the Phi Coefficient and the Chi-Square Test of Association. *psychometroscar* 2022. Available online: https://psychometroscar.com/2022/04/21/the-relationship-between-the-phi-coefficient-and-the-chi-square-test-of-association/ (accessed on 9th March 2023).
11. Cramér, H. *Mathematical Methods of Statistics*; Princeton University Press, 1999; ISBN 978-0-691-00547-8.
12. Tschuprow, A.A. *Principles of the Mathematical Theory of Correlation*; W. Hodge, limited, 1939;
13. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21*, 6, doi:10.1186/s12864-019-6413-7.
14. Rosenberg, M.S. A Generalized Formula for Converting Chi-Square Tests to Effect Sizes for Meta-Analysis. *PLoS ONE* **2010**, *5*, e10059, doi:10.1371/journal.pone.0010059.
15. Johnston, J.E.; Berry, K.J.; Mielke, P.W. Measures of Effect Size for Chi-Squared and Likelihood-Ratio Goodness-of-Fit Tests. *Percept. Mot. Skills* **2006**, *103*, 412–414, doi:10.2466/pms.103.2.412-414.