Phi, Fei, Fo, Fum: Effect Sizes for Chi-squared Tests

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 /fej/ or "fay"). We demonstrate the implementation of these measures and their confidence intervals via the `{effectsize}` package (Ben-Shachar, Lüdecke, & Makowski, 2020) in the R programming language.

Introduction

Over the last two decades, there has been growing concerns about the so-called replication crisis in psychology and other fields (Camerer et al., 2018; Open Science Collaboration, 2015). As a result, the scientific community has paid increasing attention to the issue of replicability in science, as well as to 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 (Cumming, 2014). One such recommendation coming for example from the New Statistics movement is to report effect sizes and their corresponding confidence intervals, and to increasingly rely on meta-analyses to increase confidence in those estimations. 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 Wiernik & Dahlke (2020).

Unfortunately, popular software do not always offer the necessary implementations of the specialized effect sizes necessary for a given research design. In this paper, we 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 /fej/ or "fay").

Importantly, we offer researchers an applied walkthrough on how to use these effect sizes in practice thanks to the `{effectsize}` package (Ben-Shachar et al., 2020) in the R programming language (R Core Team, 2023), which implements these measures and their confidence intervals. We cover in turn tests of independence (φ/phi, Cramér's V) and tests of goodness of fit (Cohen's w and a new proposed effect size, .(Fei/פ

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 $p = k/N$):

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

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

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

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

```
        Survived
Sex        No  Yes
  Male   1364  367
  Female  126  344
```

```
chisq.test(Titanic_xtab)
```

```

    Pearson's Chi-squared test with Yates' continuity correction

data:  Titanic_xtab
X-squared = 454.5, df = 1, p-value < 2.2e-16
```

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.

Phi

For a 2-by-2 contingency table analysis, like 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 way to compute this effect size is to re-code the binary variables as dummy (0, 1) variables, and computing the Pearson correlation between them:

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

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

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

This value ranges between 0 (no association) and 1 (complete dependence), and its values can be interpreted the same as Person's correlation coefficient.

```
library(effectsize)
library(correlation)

phi(Titanic_xtab, adjust = FALSE)
```

```
Phi  |       95% CI
-------------------
0.46 | [0.42, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

```
tidyr::uncount(as.data.frame(Titanic_xtab), weights = Freq) |>
  transform(Survived = Survived == "Yes", Sex = Sex == "Male") |>
  correlation()
```

```
# Correlation Matrix (pearson-method)

Parameter1 | Parameter2 |     r |         95% CI | t(2199) |         p
---------------------------------------------------------------------
Sex        |   Survived | -0.46 | [-0.49, -0.42] |  -24.00 | < .001***

p-value adjustment method: Holm (1979)
Observations: 2201
```

Cramér's *V*

When the contingency table is larger than 2-by-2, using $\sqrt{\chi^2/N}$ can produce values larger than 1, and so loses its interpretability as a correlation like effect size. Cramér showed (Cramér, 1999) 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 $N \times (\min(k,l) - 1)$. 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 1 when the columns are completely dependent on the rows, or the rows are completely dependent on the columns.

```
(Titanic_xtab2 <- as.table(apply(Titanic, c(1, 4), sum)))
```

```
      Survived
Class   No Yes
  1st  122 203
  2nd  167 118
  3rd  528 178
  Crew 673 212
```

```r
cramers_v(Titanic_xtab2, adjust = FALSE)
```

```
Cramer's V |       95% CI
-------------------------
0.29       | [0.26, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

Tschuprow (Tschuprow, 1939) devised an alternative value, at

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

which is 1 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.

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 if 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.

```r
data("food_class")
food_class
```

```
          Soy Milk Meat
Vegan      47    0    0
Not-Vegan   0   12   21
```

Accordingly, in such a table, Cramer's V will be 1, but Tschuprow's t will not be:

```r
cramers_v(food_class, adjust = FALSE)
```

```
Cramer's V |       95% CI
-------------------------
1.00       | [0.81, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

```r
tschuprows_t(food_class)
```

```
Tschuprow's T |       95% CI
----------------------------
0.84          | [0.68, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

We can generalize both ϕ, V, and T to: $\sqrt{\frac{\chi^2}{\chi^2_{\text{max}}}}$.

These coefficients can also be used for confusion matrices in the context of assessing machine learning algorithms classification abilities. 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 that ϕ (Chicco & Jurman, 2020).

Goodness of Fit

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

Cohen's *w*

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

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

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

```
(Titanic_freq <- as.table(apply(Titanic, 2, sum)))
```

```

  Male Female 
  1731    470 
```

```
p_E <- c(0.5, 0.5)

cohens_w(Titanic_freq, p = p_E)
```

```
Cohen's w |       95% CI
------------------------
0.57      | [0.54, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

Unfortunately, *w* has an upper bound of 1 *only* when the variable is binomial (has two categories) and the expected distribution is uniform ($p = 1 - p = 0.5$). If the distribution is none uniform (Rosenberg, 2010) or if there are more than 2 classes (Johnston, Berry, & Mielke Jr, 2006), then $\chi^2_{\text{max}} > N$, and so *w* can be larger than 1.

```
O <- c(90, 10)
p_E <- c(0.35, 0.65)
cohens_w(O, p = p_E)
```

```
Cohen's w |       95% CI
------------------------
1.15      | [0.99, 1.36]

- One-sided CIs: upper bound fixed at [1.36~].
```

```
O <- c(10, 20, 80, 5)
p_E <- c(.25, .25, .25, .25)
cohens_w(O, p = p_E)
```

```
Cohen's w |       95% CI
------------------------
1.05      | [0.88, 1.73]

- One-sided CIs: upper bound fixed at [1.73~].
```

Fei

We present here a new effect size, פ (Fei, pronounced /fej/ or “fay”), which normalizes goodness-of-fit χ^2 by the proper χ^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,

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

And so,

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

Finally, an effect size can be derived as:

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

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). פ will be 0 when the observed distribution matches the expected one perfectly, and will be 1 when the observed values are all of the same class—the one with the smallest expected probability.

```r
O <- c(90, 10)
p_E <- c(0.35, 0.65)
fei(O, p = p_E)
```

```
Fei  |       95% CI
-------------------
0.85 | [0.73, 1.00]

- Adjusted for uniform expected probabilities.
- One-sided CIs: upper bound fixed at [1.00].
```

```r
O <- c(10, 20, 80, 5)
p_E <- c(.25, .25, .25, .25)
fei(O, p = p_E)
```

```
Fei  |       95% CI
-------------------
0.60 | [0.51, 1.00]

- Adjusted for non-uniform expected probabilities.
- One-sided CIs: upper bound fixed at [1.00].
```

When there are only 2 cells with uniform expected probabilities (50%), this expression reduces to N and פ $= w$.

```r
O <- c(90, 10)
p_E <- c(0.5, 0.5)

fei(O, p = p_E)
```

```
Fei  |       95% CI
-------------------
0.80 | [0.64, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

```r
cohens_w(O, p = p_E)
```

```
Cohen's w |       95% CI
------------------------
0.80      | [0.64, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

Summary

Effect sizes are essential to interpret the magnitude of observed effects, they are frequently required in scientific journals, and they are 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 פ (Fei), we fill the missing effect size for all cases of a χ^2 test.

References

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cramér, H. (1999). *Mathematical methods of statistics* (Vol. 43). Princeton University Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

DeGeest, D. S., & Schmidt, F. L. (2010). The impact of research synthesis methods on industrial–organizational psychology: The road from pessimism to optimism about cumulative knowledge. *Research Synthesis Methods*, *1*(3-4), 185–197.

Johnston, J. E., Berry, K. J., & Mielke Jr, P. W. (2006). Measures of effect size for chi-squared and likelihood-ratio goodness-of-fit tests. *Perceptual and Motor Skills*, *103*(2), 412–414.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

R Core Team. (2023). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Rosenberg, M. S. (2010). A generalized formula for converting chi-square tests to effect sizes for meta-analysis. *PloS One*, *5*(4), e10059.

Tschuprow, A. A. (1939). *Principles of the mathematical theory of correlation*. Hodge.

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123.