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

Introduction

In both theoretical and applied research, it is often of interest to assess the strength of an observed association. Beyond their use in interpreting the results from a given study, they can be aggregated in meta-analyses (Lakens, 2013). We review here the most common effect sizes for analyses of categorical variables that use the χ^2 (chi-square) statistic, and introduce a new one—ɔ (Fei)—alongside the `{effectsize}` package (Ben-Shachar, Lüdecke, & Makowski, 2020) in the R programming language (R Core Team, 2023), which implements these measures and their confidence intervals.

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. 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.

For example, which might ask of the probability of survival of the sinking of the Titanic is dependant on the sex of the passenger:

```
(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 < 0.00000000000000022
```

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 (absolute) 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*

These properties do not hold when the contingency table is larger than 2-by-2: $\sqrt{\chi^2/N}$ can be larger than 1. 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 row are completely dependent on the columns.

```r
(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 , 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.

```r
tschuprows_t(Titanic_xtab2)
```

```
## Tschuprow's T |       95% CI
## -----------------------------
## 0.22          | [0.20, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

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

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 $\sqrt{\frac{\chi^2}{\chi^2_{\max}}}$, all we need to find is $\chi^2_{\max}$.

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

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

```
##   Male Female
##   1731    470
```

```r
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, ɒ (Fei), 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:

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

Since $\chi^2 = N \times \sum_{i=1}^{k} \frac{(p_{E_i} - p_{O_i})^2}{p_{E_i}}$, we can find $\frac{(E_i - O_i)^2}{E_i}$ for each of these values:

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

Since $\sum_{i=1}^{k} p_i = 1$, the χ^2, which is the sum of the expression above, can be retrieved as:

$$
\begin{aligned}
\chi^2_{\text{max}} &= N \times (1 - \min(p_E) + \frac{(\min(p_E) - 1)^2}{\min(p_E)}) \\
&= N \times \frac{1 - \min(p_E)}{\min(p_E)} \\
&= N \times (\frac{1}{\min(p_E)} - 1)
\end{aligned}
$$

And so 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.

```
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].
```

```
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 $= w$.

```
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].
```

```
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

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

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

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

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.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

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.