Automated Interpretation of Indices of Effect Size

Why?

The metrics used in statistics (indices of fit, model performance, or parameter estimates) can be very abstract. A long experience is required to intuitively *feel* the meaning of their values. In order to facilitate the understanding of the results they are facing, many scientists use (often implicitly) some set of **rules of thumb**. Some of these rules of thumb have been standardize and validated and subsequently published as guidelines. Understandably then, such rules of thumb are just suggestions and there is nothing universal about them. The interpretation of **any** effect size measures is always going to be relative to the discipline, the specific data, and the aims of the analyst. This is important because what might be considered a small effect in psychology might be large for some other field like public health.

One of the most famous interpretation grids was proposed by **Cohen (1988)** for a series of widely used indices, such as the correlation **r** (*r* = .20, small; *r* = .40, moderate and *r* = .60, large) or the **standardized difference** (*Cohen's d*). However, there is now a clear evidence that Cohen's guidelines (which he himself later disavowed; Funder, 2019) are much too stringent and not particularly meaningful taken out of context (Funder and Ozer 2019). This led to the emergence of a literature discussing and creating new sets of rules of thumb.

Although **everybody** agrees on the fact that effect size interpretation in a study should be justified with a rationale (and depend on the context, the field, the literature, the hypothesis, etc.), these pre-baked rules can nevertheless be useful to give a rough idea or frame of reference to understand scientific results.

The package `effectsize` catalogs such sets of rules of thumb for a variety of indices in a flexible and explicit fashion, helping you understand and report your results in a scientific yet meaningful way. Again, readers should keep in mind that these thresholds, as ubiquitous as they may be, **remain arbitrary**. Thus, their use should be discussed on a case-by-case basis depending on the field, hypotheses, prior results, and so on, to avoid their crystallization, as for the infamous $p < .05$ criterion of hypothesis testing.

Moreover, some authors suggest the counter-intuitive idea that *very large effects*, especially in the context of psychological research, is likely to be a "gross overestimate that will rarely be found in a large sample or in a replication" (Funder and Ozer 2019). They suggest that smaller effect size are worth taking seriously (as they can be potentially consequential), as well as more believable.

Correlation *r*

There can be used to interpret not only Pearson's correlation coefficient, but also Spearman's, ϕ (phi), Cramer's *V* and Tschuprow's *T*. Although Cohen's *w* and Pearson's *C* are *not* a correlation coefficients, they are often also interpreted as such.

Funder and Ozer (2019)

```
interpret_r(x, rules = "funder2019")
```

- **r < 0.05** - Tiny
- **0.05 <= r < 0.1** - Very small
- **0.1 <= r < 0.2** - Small
- **0.2 <= r < 0.3** - Medium
- **0.3 <= r < 0.4** - Large
- **r >= 0.4** - Very large

Gignac and Szodorai (2016)

Gignac's rules of thumb are actually one of few interpretation grid justified and based on actual data, in this case on the distribution of effect magnitudes in the literature.

```
interpret_r(x, rules = "gignac2016")
```

- **r < 0.1** - Very small
- **0.1 <= r < 0.2** - Small
- **0.2 <= r < 0.3** - Moderate
- **r >= 0.3** - Large

J. Cohen (1988)

```
interpret_r(x, rules = "cohen1988")
```

- **r < 0.1** - Very small
- **0.1 <= r < 0.3** - Small
- **0.3 <= r < 0.5** - Moderate
- **r >= 0.5** - Large

Evans (1996)

```
interpret_r(x, rules = "evans1996")
```

- **r < 0.2** - Very weak
- **0.2 <= r < 0.4** - Weak

- **0.4 <= r < 0.6** - Moderate
  - **0.6 <= r < 0.8** - Strong
  - **r >= 0.8** - Very strong

Lovakov and Agadullina (2021)

```
interpret_r(x, rules = "lovakov2021")
```

  - **r < 0.12** - Very small
  - **0.12 <= r < 0.24** - Small
  - **0.24 <= r < 0.41** - Moderate
  - **r >= 0.41** - Large

Standardized Difference *d* (Cohen's *d*)

The standardized difference can be obtained through the standardization of linear model's parameters or data, in which they can be used as indices of effect size.

J. Cohen (1988)

```
interpret_cohens_d(x, rules = "cohen1988")
```

  - **d < 0.2** - Very small
  - **0.2 <= d < 0.5** - Small
  - **0.5 <= d < 0.8** - Medium
  - **d >= 0.8** - Large

Sawilowsky (2009)

```
interpret_cohens_d(x, rules = "sawilowsky2009")
```

  - **d < 0.1** - Tiny
  - **0.1 <= d < 0.2** - Very small
  - **0.2 <= d < 0.5** - Small
  - **0.5 <= d < 0.8** - Medium
  - **0.8 <= d < 1.2** - Large
  - **1.2 <= d < 2** - Very large
  - **d >= 2** - Huge

Gignac and Szodorai (2016)

Gignac's rules of thumb are actually one of few interpretation grid justified and based on actual data, in this case on the distribution of effect magnitudes in the literature. These is in fact the same grid used for *r*, based on the conversion of *r* to *d*:

```
interpret_cohens_d(x, rules = "gignac2016")
```

- **d < 0.2** - Very small
- **0.2 <= d < 0.41** - Small
- **0.41 <= d < 0.63** - Moderate
- **d >= 0.63** - Large

Lovakov and Agadullina (2021)

```
interpret_cohens_d(x, rules = "lovakov2021")
```

- **r < 0.15** - Very small
- **0.15 <= r < 0.36** - Small
- **0.36 <= r < 0.65** - Moderate
- **r >= 0.65** - Large

Odds Ratio (OR)

Odds ratio, and *log* odds ratio, are often found in epidemiological studies. However, they are also the parameters of *logistic* regressions, where they can be used as indices of effect size. Note that the (log) odds ratio from logistic regression coefficients are *unstandardized*, as they depend on the scale of the predictor. In order to apply the following guidelines, make sure you *standardize* your predictors!

Keep in mind that these apply to Odds *ratios*, so Odds ratio of 10 is as extreme as a Odds ratio of 0.1 (1/10).

Chen, Cohen, and Chen (2010)

```
interpret_oddsratio(x, rules = "chen2010")
```

- **OR < 1.68** - Very small
- **1.68 <= OR < 3.47** - Small
- **3.47 <= OR < 6.71** - Medium
- **OR >= 6.71** - Large

J. Cohen (1988)

```
interpret_oddsratio(x, rules = "cohen1988")
```

- **OR < 1.44** - Very small
- **1.44 <= OR < 2.48** - Small
- **2.48 <= OR < 4.27** - Medium
- **OR >= 4.27** - Large

This converts (log) odds ratio to standardized difference *d* using the following formula (J. Cohen 1988; Sánchez-Meca, Marín-Martínez, and Chacón-Moscoso 2003):

$$d = log(OR) \times \frac{\sqrt{3}}{\pi}$$

Coefficient of determination (R^2)

For Linear Regression

J. Cohen (1988)

```
interpret_r2(x, rules = "cohen1988")
```

- **R2 < 0.02** - Very weak
- **0.02 <= R2 < 0.13** - Weak
- **0.13 <= R2 < 0.26** - Moderate
- **R2 >= 0.26** - Substantial

Falk and Miller (1992)

```
interpret_r2(x, rules = "falk1992")
```

- **R2 < 0.1** - Negligible
- **R2 >= 0.1** - Adequate

For PLS / SEM R-Squared of *latent* variables

Chin et al. (1998)

```
interpret_r2(x, rules = "chin1998")
```

- **R2 < 0.19** - Very weak
- **0.19 <= R2 < 0.33** - Weak
- **0.33 <= R2 < 0.67** - Moderate
- **R2 >= 0.67** - Substantial

Hair, Ringle, and Sarstedt (2011)

```
interpret_r2(x, rules = "hair2011")
```

- **R2 < 0.25** - Very weak
- **0.25 <= R2 < 0.50** - Weak
- **0.50 <= R2 < 0.75** - Moderate
- **R2 >= 0.75** - Substantial

Omega / Eta / Epsilon Squared

The Omega squared is a measure of effect size used in ANOVAs. It is an estimate of how much variance in the response variables are accounted for by the explanatory variables. Omega squared is widely viewed as a lesser biased alternative to eta-squared, especially when sample sizes are small.

Field (2013)

```
interpret_omega_squared(x, rules = "field2013")
```

- **ES < 0.01** - Very small
- **0.01 <= ES < 0.06** - Small
- **0.06 <= ES < 0.14** - Medium
- **ES >= 0.14** - Large

Jacob Cohen (1992)

These are applicable to one-way ANOVAs, or to *partial* Eta / Omega / Epsilon Squared in a multi-way ANOVA.

```
interpret_omega_squared(x, rules = "cohen1992")
```

- **ES < 0.02** - Very small
- **0.02 <= ES < 0.13** - Small
- **0.13 <= ES < 0.26** - Medium
- **ES >= 0.26** - Large

Kendall's coefficient of concordance

The interpretation of Kendall's coefficient of concordance (*w*) is a measure of effect size used in non-parametric ANOVAs (the Friedman rank sum test). It is an estimate of agreement among multiple raters.

Landis and Koch (1977)

```
interpret_omega_squared(w, rules = "landis1977")
```

- **0.00 <= w < 0.20** - Slight agreement
- **0.20 <= w < 0.40** - Fair agreement
- **0.40 <= w < 0.60** - Moderate agreement
- **0.60 <= w < 0.80** - Substantial agreement
- **w >= 0.80** - Almost perfect agreement

Cohen's *g*

Cohen's *g* is a measure of effect size used for McNemar's test of agreement in selection - when

repeating a multiple chose selection, is the percent of matches (first response is equal to the second response) different than 50%?

J. Cohen (1988)

```
interpret_cohens_g(x, rules = "cohen1988")
```

- **d < 0.05** - Very small
- **0.05 <= d < 0.15** - Small
- **0.15 <= d < 0.25** - Medium
- **d >= 0.25** - Large

Interpretation of other Indices

`effectsize` also offers functions for interpreting other statistical indices:

- `interpret_gfi()`, `interpret_agfi()`, `interpret_nfi()`, `interpret_nnfi()`, `interpret_cfi()`, `interpret_rmsea()`, `interpret_srmr()`, `interpret_rfi()`, `interpret_ifi()`, and `interpret_pnfi()` for interpretation CFA / SEM goodness of fit.
- `interpret_p()` for interpretation of *p*-values.
- `interpret_direction()` for interpretation of direction.
- `interpret_bf()` for interpretation of Bayes factors.
- `interpret_rope()` for interpretation of Bayesian ROPE tests.
- `interpret_ess()` and `interpret_rhat()` for interpretation of Bayesian diagnostic indices.

References

Chen, Henian, Patricia Cohen, and Sophie Chen. 2010. "How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies." *Communications in Statistics—Simulation and Computation* 39 (4): 860–64.

Chin, Wynne W et al. 1998. "The Partial Least Squares Approach to Structural Equation Modeling." *Modern Methods for Business Research* 295 (2): 295–336.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences, 2nd Ed.* New York: Routledge.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155.

Evans, James D. 1996. *Straightforward Statistics for the Behavioral Sciences.* Thomson Brooks/Cole Publishing Co.

Falk, R Frank, and Nancy B Miller. 1992. *A Primer for Soft Modeling.* University of Akron Press.

Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics*. sage.

Funder, David C, and Daniel J Ozer. 2019. "Evaluating Effect Size in Psychological Research: Sense and Nonsense." *Advances in Methods and Practices in Psychological Science*, 2515245919847202.

Gignac, Gilles E, and Eva T Szodorai. 2016. "Effect Size Guidelines for Individual Differences Researchers." *Personality and Individual Differences* 102: 74–78.

Hair, Joe F, Christian M Ringle, and Marko Sarstedt. 2011. "PLS-SEM: Indeed a Silver Bullet." *Journal of Marketing Theory and Practice* 19 (2): 139–52.

Landis, J Richard, and Gary G Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, 159–74.

Lovakov, Andrey, and Elena R Agadullina. 2021. "Empirically Derived Guidelines for Effect Size Interpretation in Social Psychology." *European Journal of Social Psychology*.

Sánchez-Meca, Julio, Fulgencio Marín-Martínez, and Salvador Chacón-Moscoso. 2003. "Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis." *Psychological Methods* 8 (4): 448.

Sawilowsky, Shlomo S. 2009. "New Effect Size Rules of Thumb."