Effect-size formulas

Purpose

`testflow` reports one primary effect-size estimate with each workflow when an effect size is defined. The formulas below describe the implemented estimators. They are written to match the package code, including the direction of signed effects.

Cohen's d

One-sample Cohen's d

For a numeric sample $x_1, \ldots, x_n$ compared with reference value μ_0, `test_one_sample()` reports:

$$d = \frac{\bar{x} - \mu_0}{s_x}$$

where $\bar{x}$ is the sample mean and s_x is the sample standard deviation, both computed after removing missing values.

Independent-groups Cohen's d

For two independent groups x and y, `test_two_groups()` reports Cohen's d on the Student and Welch t-test branches:

$$d = \frac{\bar{x} - \bar{y}}{s_p}$$

with pooled standard deviation:

$$s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

The sign follows the group order used internally by the workflow: the first group minus the second group.

Paired-sample Cohen's dz

For paired measurements, `test_paired()` computes each paired difference as:

$$d_i = after_i - before_i$$

and reports Cohen's d_z:

$$d_z = \frac{\bar{d}}{s_d}$$

where $\bar{d}$ and s_d are the mean and standard deviation of the paired differences.

ANOVA-style effect sizes

Eta squared

For one-way and factorial ANOVA workflows, `testflow` reports eta squared for the selected ANOVA term:

$$\eta_j^2 = \frac{SS_j}{\sum SS}$$

where SS_j is the term sum of squares and $\sum SS$ is the sum of all ANOVA-table sums of squares, including residual variation.

For repeated-measures ANOVA, the implemented repeated-time eta squared is:

$$\eta_{time}^2 = \frac{SS_{time}}{SS_{total}}$$

with:

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$

and:

$$SS_{time} = \sum_t n_t (\bar{y}_t - \bar{y})^2$$

where n_t is the number of observations at time t, $\bar{y}_t$ is the mean at time t, and $\bar{y}$ is the grand mean.

Kruskal-Wallis epsilon squared

For the non-parametric branch of `test_groups()`, the reported effect is:

$$\epsilon^2 = \frac{H - k + 1}{n - k}$$

where H is the Kruskal-Wallis statistic, k is the number of groups, and n is the number of complete observations.

Regression effect sizes

R squared and adjusted R squared

For `test_linear_regression()`:

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}, \qquad R^2_{adj} = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

where p is the number of predictors and n is the number of complete observations. Both are reported, with `R squared` first (used as the primary effect size) and `Adjusted R squared` second.

McFadden's pseudo R squared

For `test_logistic_regression()`:

$$R^2_{McFadden} = 1 - \frac{\log L_{model}}{\log L_{null}}$$

where L_{model} is the fitted model's likelihood and L_{null} is the likelihood of the intercept-only model.

Survival effect sizes

Hazard ratio

For `test_survival()`, the effect size is the hazard ratio from a companion univariate Cox model on the same grouping factor used by the log-rank test:

$$HR = e^{\hat{\beta}}$$

reported with its Wald confidence interval. The hazard ratio is not part of the log-rank test itself and is not assigned a magnitude label: unlike Cohen's d or R^2, there is no widely agreed small/moderate/large convention for hazard ratios, and inventing one would overstate how standardized that judgment is. The estimate is always reported; only the magnitude label is omitted.

Concordance index

For `test_cox()`, the effect size is Harrell's concordance index (C), the proportion of comparable subject pairs for which the model correctly orders predicted risk and observed survival time. Like the hazard ratio, it is reported without a magnitude label for the same reason: 0.5 corresponds to no better than chance and 1.0 to perfect discrimination, but there is no standard small/moderate/large convention in between.

Categorical and rank-based effect sizes

Cramer's V

For categorical association workflows, `testflow` reports:

$$V = \sqrt{\frac{\chi^2}{n(\min(r, c) - 1)}}$$

where χ^2 is the Pearson chi-square statistic, n is the table total, r is the number of rows, and c is the number of columns.

Rank-biserial correlation

For the Wilcoxon branch of `test_two_groups()`, the reported rank-biserial correlation is:

$$r_{rb} = 1 - \frac{2W}{n_1 n_2}$$

where W is the `stats::wilcox.test()` statistic for the first group after R's conversion to the Mann-Whitney U scale, and n_1 and n_2 are the non-missing group sizes. The sign follows the same first-group reference.

Kendall's W

For Friedman repeated numeric workflows, `testflow` reports:

$$W = \frac{\chi_F^2}{n(k-1)}$$

where χ_F^2 is the Friedman statistic, n is the number of complete subjects, and k is the number of repeated measurements.

For Cochran Q repeated categorical workflows, the implemented analogous effect is:

$$W = \frac{Q}{n(k-1)}$$

where Q is the Cochran Q statistic.

Diagnostic and agreement effect sizes

Accuracy

For `test_diagnostic()`, the effect size is overall accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

No magnitude label is assigned; accuracy is only meaningful relative to the no-information rate reported alongside it (the primary test result).

AUC

For `test_roc()`, the effect size is the area under the ROC curve (see "Diagnostic and agreement effect sizes" formulas in the function documentation for the full derivation). The magnitude label follows Hosmer, Lemeshow & Sturdivant (2013, p. 177): poor if < 0.7, acceptable if < 0.8, excellent if < 0.9, otherwise outstanding (applied to $\max(AUC, 1 - AUC)$, since an AUC below 0.5 indicates the predictor discriminates in the opposite direction, not that it fails to discriminate).

Cohen's kappa

For `test_agreement()`, the effect size is Cohen's kappa. The magnitude label follows Landis & Koch (1977): poor if < 0, slight if < 0.21, fair if < 0.41, moderate if < 0.61, substantial if < 0.81, otherwise almost perfect.

Intraclass correlation coefficient

For `test_icc()`, the effect size is ICC(2,1) (two-way random effects, absolute agreement). The magnitude label follows Koo & Li (2016): poor if < 0.5, moderate if < 0.75, good if < 0.9, otherwise excellent.

Other reported workflow summaries

Some workflows expose a scalar summary in the effect-size field when a standard Cohen-style effect size is not used:

- `test_proportion()` reports the observed success proportion:

$$\hat{p} = \frac{x}{n}$$

- `test_paired_categorical()` reports the number of discordant pairs:

$$b + c$$

- `test_multinomial()` reports the chi-square goodness-of-fit statistic:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- `test_outliers()` reports the number of rows flagged by the selected outlier rule.

Magnitude labels

Magnitude labels are descriptive thresholds used consistently by the package:

- Cohen's d: negligible if $|d| < 0.2$, small if $|d| < 0.5$, moderate if $|d| < 0.8$, otherwise large.
- Eta squared, epsilon squared, R squared, adjusted R squared, and McFadden's pseudo R squared: negligible if the estimate is < 0.01, small if < 0.06, moderate if < 0.14, otherwise large. (McFadden's pseudo R squared is not on the same scale as ordinary least squares R^2; values above roughly 0.2-0.4 are already considered a good fit in practice, so treat the "large" label loosely for this estimate.)
- Cramer's V, rank-biserial absolute value, and Kendall's W: negligible if < 0.1, small if < 0.3, moderate if < 0.5, otherwise large.
- Hazard ratio and concordance index: no magnitude label is assigned (see "Survival effect sizes" above); the estimate itself is always reported.
- Accuracy: no magnitude label is assigned (see "Diagnostic and agreement effect sizes" above); only meaningful relative to the no-information rate.
- AUC: poor if $\max(AUC, 1 - AUC) < 0.7$, acceptable if < 0.8, excellent if < 0.9, otherwise outstanding.
- Cohen's kappa: poor if < 0, slight if < 0.21, fair if < 0.41, moderate if < 0.61, substantial if < 0.81, otherwise almost perfect.
- Intraclass correlation coefficient: poor if < 0.5, moderate if < 0.75, good if < 0.9, otherwise excellent.