

Statistical test workflows

Introduction

testflow provides statistical testing workflows organized by study design.

One numerical variable

```
library(testflow)
cardio <- make_cardio_data()
test_one_sample(cardio, sbp_3m, mu = 140)
#> One-Sample Test
#>
#> Outcome: sbp_3m
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality: sbp_3m: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk; stati
#> * Symmetry of deviations: acceptable: The deviations from the reference value look approximately sym
#>
#> Recommended test
#> One-sample t-test
#>
#> Result
#> H0: the population mean or location of sbp_3m equals the reference value.
#> statistic = -0.46, df = 179.00, p = 0.646, 95% CI [136.47, 142.20]
#>
#> Effect size
#> Cohen's d: -0.03, negligible
#>
#> Report
#> The one numerical sample workflow for sbp_3m did not show a statistically significant result using O
```

Two independent groups

```
test_two_groups(sbp_3m ~ sex, data = cardio)
#> Two Independent Groups
#>
#> Outcome: sbp_3m
#> Group: sex
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality: sbp_3m (female): acceptable: Approximate normality looks reasonable. (method=Shapiro-Wi
#> * Normality: sbp_3m (male): acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable. (method=Levene test; stat
#> * Extreme outliers: warning: 4 potential outlier(s) flagged by IQR. (IQR rule, n = 4)
#> * Variance ratio check: acceptable: Variance ratio looks reasonable. (statistic=1.27)
```

```

#>
#> Recommended test
#> Student independent t-test
#>
#> Result
#> H0: the population mean or location of sbp_3m is equal across levels of sex.
#> statistic = -1.91, df = 178.00, p = 0.058, 95% CI [-11.22, 0.18]
#>
#> Effect size
#> Cohen's d: -0.29, small
#>
#> Report
#> The two independent groups workflow for sbp_3m did not show a statistically significant result using

```

Paired measurements

```

test_paired(sbp_3m ~ sbp_baseline, data = cardio)
#> Paired Measurements
#>
#> Outcome: sbp_3m - sbp_baseline
#>
#> Assumptions
#> * Independence of observations: assumed: Paired observations from the same subjects are assumed by d
#> * Normality: diff: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk; statist
#> * Symmetry of paired differences: acceptable: The deviations from the reference value look approxima
#> * Extreme outliers: warning: 1 potential outlier(s) flagged by IQR. (IQR rule, n = 1)
#>
#> Recommended test
#> Paired t-test
#>
#> Result
#> H0: the mean or median paired difference (sbp_3m - sbp_baseline) equals 0.
#> statistic = -9.20, df = 179.00, p = <0.001, 95% CI [-9.53, -6.16]
#>
#> Effect size
#> Cohen's dz: -0.69, moderate
#>
#> Report
#> The paired measurements workflow for sbp_3m - sbp_baseline showed a statistically significant result

```

More than two groups

```

test_groups(sbp_3m ~ treatment, data = cardio)
#> More Than Two Independent Groups
#>
#> Outcome: sbp_3m
#> Group: treatment
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality: sbp_3m (lifestyle): not acceptable: Normality may be violated. (method=Shapiro-Wilk; st
#> * Normality: sbp_3m (medication): acceptable: Approximate normality looks reasonable. (method=Shapir

```

```

#> * Normality: sbp_3m (usual care): acceptable: Approximate normality looks reasonable. (method=Shapir
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable. (method=Levene test; stat
#> * Bartlett test: acceptable: Variance homogeneity looks reasonable. (method=Bartlett test; statistic
#> * Extreme outliers: warning: 4 potential outlier(s) flagged by IQR. (IQR rule, n = 4)
#>
#> Recommended test
#> Kruskal-Wallis test
#>
#> Result
#> H0: the population mean or location of sbp_3m is equal across levels of treatment.
#> statistic = 7.58, df = 2.00, p = 0.023
#>
#> Effect size
#> Kruskal epsilon squared: 0.03, small
#>
#> Report
#> The more than two independent groups workflow for sbp_3m showed a statistically significant result u

```

Factorial designs

```

test_factorial(sbp_3m ~ sex * treatment, data = cardio)
#> Factorial ANOVA
#>
#> Outcome: sbp_3m
#> Group: sex, treatment
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality of residuals: acceptable: Residuals appear approximately normal. (method=Shapiro-Wilk; s
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable. (method=Levene test; stat
#> * Balanced design: not required: Cell sizes are unbalanced; Type II sums of squares are used so the
#>
#> Recommended test
#> Factorial ANOVA
#>
#> Result
#> H0: the population mean or location of sbp_3m is equal across levels of sex, treatment.
#> statistic = 5.01, df = 1.00, p = 0.027
#>
#> Effect size
#> eta squared: 0.03, small
#>
#> ANOVA table
#>
#>      term      sumsq      df statistic p.value
#>      sex  1808.24      1.00      5.01  0.027
#>      treatment 3601.18      2.00      4.99  0.008
#>      sex:treatment   97.59      2.00      0.14  0.874
#>      Residuals 62839.72 174.00      <NA>    <NA>
#>
#> Report
#> The factorial design workflow for sbp_3m showed a statistically significant result using Factorial A

```

Repeated measurements

```
test_repeated(cardio, c(sbp_baseline, sbp_3m, sbp_6m), id = id)
#> Repeated Measures
#>
#> Outcome: sbp_baseline, sbp_3m, sbp_6m
#> Group: time
#>
#> Assumptions
#> * Independence of observations: assumed: Repeated measurements from the same subjects are assumed by
#> * Normality: sbp_3m: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk; stati
#> * Normality: sbp_6m: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk; stati
#> * Normality: sbp_baseline: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk;
#> * Sphericity: warning: Sphericity may be violated; interpret the repeated-measures ANOVA with caution
#>
#> Recommended test
#> Repeated-measures ANOVA
#>
#> Result
#> H0: the population mean or location of sbp_baseline, sbp_3m, sbp_6m is equal across levels of time.
#> statistic = 68.14, df = 2.00, p = <0.001
#>
#> Effect size
#> eta squared: 0.05, small
#>
#> Post-hoc comparisons
#>
#>      group1 group2      method statistic      p  p.adj p.adjust.method
#> sbp_baseline sbp_3m paired t-test      -9.20 <0.001 <0.001             BH
#> sbp_baseline sbp_6m paired t-test      -8.89 <0.001 <0.001             BH
#>      sbp_3m sbp_6m paired t-test      -3.54 <0.001 <0.001             BH
#>
#> Report
#> The repeated numeric measurements workflow for sbp_baseline, sbp_3m, sbp_6m showed a statistically s
```

The repeated numeric workflow chooses repeated-measures ANOVA when the within-time normality checks are acceptable and Friedman otherwise. Post-hoc comparisons are paired t-tests for the parametric branch and paired Wilcoxon tests for the non-parametric branch.

Categorical outcomes

```
test_categorical(treatment ~ controlled_3m, data = cardio)
#> Categorical Association
#>
#> Outcome: treatment
#> Group: controlled_3m
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Expected cell counts: acceptable: Chi-square approximation is reasonable. (method=Pearson chi-squa
#>
#> Recommended test
#> Chi-square test of independence
#>
#> Result
```

```
#> H0: treatment and controlled_3m are independent.
#> statistic = 5.02, df = 2.00, p = 0.081
#>
#> Effect size
#> Cramer's V: 0.17, small
#>
#> Report
#> The two categorical variables workflow for treatment did not show a statistically significant result
```

Repeated categorical outcomes

```
test_repeated_categorical(cardio, c(controlled_baseline, controlled_3m, controlled_6m))
#> Repeated Categorical Measurements
#>
#> Outcome: controlled_baseline, controlled_3m, controlled_6m
#>
#> Assumptions
#> * Repeated binary measurements: assumed: Same subjects should be measured at 3 or more time points.
#> * Complete repeated data: acceptable: Missingness should be handled explicitly or via complete-case
#>
#> Recommended test
#> Cochran Q test
#>
#> Result
#> H0: the success proportions are equal across repeated categorical measures.
#> statistic = 39.58, df = 2.00, p = <0.001
#>
#> Effect size
#> Cochran Q Kendall's W: 0.11, small
#>
#> Pairwise McNemar
#>
#>      group1      group2      method statistic      p      p.adj p.adjust.method
#> controlled_baseline controlled_3m McNemar test      18.37 <0.001 <0.001              BH
#> controlled_baseline controlled_6m McNemar test      25.35 <0.001 <0.001              BH
#>      controlled_3m controlled_6m McNemar test       1.71  0.190  0.190              BH
#>
#> Report
#> The repeated categorical measurements workflow for controlled_baseline, controlled_3m, controlled_6m
```

The repeated categorical workflow uses Cochran Q for binary repeated outcomes and pairwise McNemar tests for follow-up comparisons.

References

- Fisher, R. A. (1925). *Statistical Methods for Research Workers*.
- Gosset, W. S. (1908). The probable error of a mean.
- Welch, B. L. (1947). Generalization of Student's problem with unequal variances.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other.
- Levene, H. (1960). Robust tests for equality of variances.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance.

- Dunn, O. J. (1964). Multiple comparisons using rank sums.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance.
- Cochran, W. G. (1950). The comparison of percentages in matched samples.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages.
- Pearson, K. (1895, 1900).
- Spearman, C. (1904). The proof and measurement of association between two things.
- Kendall, M. G. (1938). A new measure of rank correlation.
- Cramer, H. (1946). *Mathematical Methods of Statistics*.
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*.
- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.).
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.).
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration.
- Cox, D. R. (1972). Regression models and life-tables.
- Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals.
- Simel, D. L., Samsa, G. P., & Matchar, D. B. (1991). Likelihood ratios with confidence: sample size estimation for diagnostic test studies.
- Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: sensitivity and specificity.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.
- Youden, W. J. (1950). Index for rating diagnostic tests.
- Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research.

Correlation

```
test_correlation(sbp_3m ~ age, data = cardio)
#> Correlation
#>
#> Outcome: sbp_3m
#> Group: age
#>
#> Assumptions
#> * Monotonic relationship: warning: Relationship may be non-monotonic. (method=Spearman correlation;
#> * Extreme outliers: warning: 7 potential outlier(s) flagged by IQR. (IQR rule applied to age, sbp_3m
#> * Normality: not required: Normality is not required for Spearman correlation.
#>
#> Recommended test
#> Spearman Correlation
#>
#> Result
#> H0: the correlation between age and sbp_3m is 0.
```

```

#> statistic = 793638.65, p = 0.014
#>
#> Effect size
#> Spearman Correlation r: 0.18, small
#>
#> Method comparison
#>   method statistic p.value
#>   Pearson      0.15  0.041
#>   Spearman      0.18  0.014
#>   Kendall      0.12  0.016
#>
#> Report
#> The two numeric variables workflow for sbp_3m showed a statistically significant result using Spearman

```

Regression

```

test_linear_regression(sbp_3m ~ age + ldl, data = cardio)
#> Linear Regression
#>
#> Outcome: sbp_3m
#> Predictors: age, ldl
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality of residuals: acceptable: Residuals appear approximately normal. (method=Shapiro-Wilk; s
#> * Homoscedasticity: acceptable: Residual variance looks constant across fitted values. (method=Breus
#> * Multicollinearity: acceptable: Variance inflation factors are all below 5. (method=Variance inflat
#>
#> Recommended test
#> Linear regression
#>
#> Result
#> H0: none of age, ldl explain variation in sbp_3m (all slopes equal 0).
#> statistic = 2.91, df = 2.00, p = 0.057
#>
#> Effect size
#> R squared: 0.03, small
#>
#> Coefficients
#>      term estimate std.error statistic p.value conf.low conf.high
#> (Intercept)  115.07    10.27     11.20 <0.001    94.79    135.34
#>      age      0.29     0.13      2.16  0.032     0.03     0.56
#>      ldl      2.28     1.81      1.26  0.209    -1.29     5.86
#>
#> Report
#> The multiple linear regression workflow for sbp_3m did not show a statistically significant result u

```

Linear regression reports the overall model F-test as the primary result, the per-term coefficient table (with confidence intervals) in `alternative_tests$coefficients`, and R squared / adjusted R squared as the effect size. Residual normality (Shapiro-Wilk), homoscedasticity (Breusch-Pagan), and multicollinearity (variance inflation factor, when there is more than one predictor) are checked as assumptions.

```

test_logistic_regression(controlled_3m ~ age + ldl, data = cardio)
#> Logistic Regression
#>
#> Outcome: controlled_3m
#> Predictors: age, ldl
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Multicollinearity: acceptable: Variance inflation factors are all below 5. (method=Variance inflation factors)
#> * Influential observations: warning: 5 observation(s) exceed the Cook's distance threshold (4/n); consider removing
#>
#> Recommended test
#> Logistic regression
#>
#> Result
#> H0: none of age, ldl are associated with controlled_3m (all log-odds slopes equal 0).
#> statistic = 4.65, df = 2.00, p = 0.098
#>
#> Effect size
#> McFadden's pseudo R squared: 0.02, small
#>
#> Odds ratios
#>
#>      term odds_ratio std.error statistic p.value or.conf.low or.conf.high
#> (Intercept)      9.61      1.10      2.05  0.040      1.15      87.80
#>      age         0.97      0.01     -2.03  0.042      0.94      1.00
#>      ldl         0.86      0.19     -0.79  0.431      0.59      1.25
#>
#> Report
#> The logistic regression workflow for controlled_3m did not show a statistically significant result u

```

Logistic regression reports the likelihood-ratio test against the intercept-only model as the primary result, coefficients on both the log-odds and odds-ratio scale in `alternative_tests`, and McFadden's pseudo R squared as the effect size. Multicollinearity and influential observations (Cook's distance) are checked as assumptions.

Survival analysis

`make_cardio_data()` does not include time-to-event data, so this section uses a small synthetic example instead.

```

set.seed(1)
n <- 100
survival_dat <- tibble::tibble(
  time = rexp(n, 0.1),
  status = rbinom(n, 1, 0.7),
  arm = rep(c("control", "treatment"), each = n / 2),
  age = rnorm(n, 60, 10)
)
test_survival(Surv(time, status) ~ arm, data = survival_dat)
#> Kaplan-Meier / Log-Rank
#>
#> Time: time
#> Group: arm
#>

```

```

#> Assumptions
#> * Independence of observations: assumed: Independent censoring is assumed.
#>
#> Recommended test
#> Log-rank test
#>
#> Result
#> H0: the survival distributions are equal across levels of arm.
#> statistic = 0.03, df = 1.00, p = 0.861, 95% CI [0.66, 1.65]
#>
#> Effect size
#> Hazard ratio: 1.04
#>
#> Companion hazard ratio
#>      term estimate std.error statistic p.value conf.low conf.high
#> armtreatment    1.04     0.23     0.18  0.861    0.66    1.65
#>
#> Report
#> The Kaplan-Meier and log-rank workflow for time did not show a statistically significant result using

```

`test_survival()` reports the log-rank test as the primary result and a companion univariate Cox hazard ratio as the effect size; the hazard ratio is not itself part of the log-rank test and can disagree with it under strongly non-proportional hazards. `Surv()` is re-exported from the `survival` package, so it is available after `library(testflow)` alone.

```

test_cox(Surv(time, status) ~ age + arm, data = survival_dat)
#> Cox Proportional Hazards Regression
#>
#> Time: time
#> Predictors: age, arm
#>
#> Assumptions
#> * Independence of observations: assumed: Independent censoring is assumed.
#> * Proportional hazards: acceptable: No evidence against proportional hazards. (method=Schoenfeld res
#>
#> Recommended test
#> Cox proportional hazards regression
#>
#> Result
#> H0: none of age, arm are associated with the hazard (all log hazard ratios equal 0).
#> statistic = 1.42, df = 2.00, p = 0.491
#>
#> Effect size
#> Concordance index: 0.55
#>
#> Hazard ratios
#>      term hazard_ratio std.error statistic p.value conf.low conf.high
#>      age           0.99     0.01    -1.17  0.241    0.96    1.01
#> armtreatment      1.06     0.24     0.24  0.807    0.67    1.68
#>
#> Report
#> The Cox proportional hazards regression workflow for time did not show a statistically significant r

```

`test_cox()` reports the overall likelihood-ratio test as the primary result, hazard ratios per term, the

concordance index as the effect size, and checks the proportional-hazards assumption via the Schoenfeld residual test (`survival::cox.zph()`).

Diagnostic test accuracy and agreement

`make_cardio_data()` has no diagnostic-test or multi-rater columns, so this section uses small synthetic examples instead.

```
set.seed(1)
diag_dat <- tibble::tibble(
  test = c(rep("positive", 55), rep("negative", 98)),
  reference = c(rep("positive", 45), rep("negative", 10), rep("positive", 8), rep("negative", 90))
)
test_diagnostic(diag_dat, test, reference)
#> Diagnostic Test Accuracy
#>
#> Test: test
#> Reference: reference
#>
#> Assumptions
#> * Independence of observations: assumed: Test and reference results are assumed to be measured indep
#>
#> Recommended test
#> Diagnostic accuracy (2x2 table)
#>
#> Result
#> H0: accuracy equals the no-information rate (0.65).
#> statistic = 135.00, df = 153.00, p = <0.001, 95% CI [0.83, 1.00]
#>
#> Effect size
#> Accuracy: 0.88
#>
#> Diagnostic accuracy
#>
#>          metric estimate conf.low conf.high
#>          Sensitivity    0.85    0.72    0.93
#>          Specificity    0.90    0.82    0.95
#> Positive predictive value 0.82    0.69    0.91
#> Negative predictive value 0.92    0.85    0.96
#>          Accuracy    0.88    0.82    0.93
#> Positive likelihood ratio  8.49    4.67   15.45
#> Negative likelihood ratio  0.17    0.09    0.32
#>
#> Report
#> The diagnostic test accuracy workflow for test showed a statistically significant result using Diagn
```

`test_diagnostic()` reports sensitivity, specificity, predictive values, and likelihood ratios (each with an exact confidence interval) in `alternative_tests$diagnostic_table`, and tests overall accuracy against the no-information rate as the primary result, following the same convention as `caret::confusionMatrix()`.

```
roc_dat <- tibble::tibble(
  marker = c(rnorm(60, 2, 1), rnorm(50, 0, 1)),
  disease = c(rep("yes", 60), rep("no", 50))
)
test_roc(roc_dat, marker, disease)
#> ROC Curve
```

```

#>
#> Predictor: marker
#> Outcome: disease
#>
#> Assumptions
#> * Independence of observations: assumed: Higher predictor values are assumed to indicate the positive outcome
#>
#> Recommended test
#> ROC / AUC analysis
#>
#> Result
#> H0: AUC = 0.5 (marker does not discriminate between levels of disease).
#> statistic = 17.24, p = <0.001, 95% CI [0.88, 0.98]
#>
#> Effect size
#> AUC: 0.93, outstanding
#>
#> Optimal threshold (Youden's J)
#> threshold sensitivity specificity youden_j
#>      1.29      0.87      0.88      0.75
#>
#> Report
#> The ROC curve workflow for marker showed a statistically significant result using ROC / AUC analysis

```

`test_roc()` computes the AUC from the Mann-Whitney relationship, a closed-form Hanley-McNeil confidence interval, and the Youden's-J-optimal threshold.

```

agree_dat <- tibble::tibble(
  rater1 = sample(c("mild", "moderate", "severe"), 100, replace = TRUE),
  rater2 = sample(c("mild", "moderate", "severe"), 100, replace = TRUE)
)
test_agreement(agree_dat, rater1, rater2)
#> Inter-Rater Agreement
#>
#> Rater 1: rater1
#> Rater 2: rater2
#>
#> Assumptions
#> * Independence of observations: assumed: Raters are assumed to classify subjects independently.
#>
#> Recommended test
#> Cohen's kappa
#>
#> Result
#> H0: agreement beyond chance is zero (kappa = 0).
#> statistic = 1.16, p = 0.248, 95% CI [-0.06, 0.22]
#>
#> Effect size
#> Cohen's kappa: 0.08, slight
#>
#> Agreement table
#>   Rater 1 Rater 2 n
#>   mild    mild 18
#> moderate mild 12

```

```
#>   severe      mild 8
#>      mild moderate 9
#> moderate moderate 12
#>   severe moderate 5
#>      mild      severe 17
#> moderate      severe 10
#>   severe      severe 9
#>
#> Report
#> The inter-rater agreement workflow for rater1 did not show a statistically significant result using
```

`test_agreement()` reports Cohen's kappa with the Fleiss-Cohen-Everitt (1969) large-sample confidence interval.

```
icc_dat <- tibble::tibble(
  rater1 = rnorm(30, 50, 10),
  rater2 = rnorm(30, 50, 10),
  rater3 = rnorm(30, 50, 10)
)
test_icc(icc_dat, c(rater1, rater2, rater3))
#> Intraclass Correlation
#>
#> Outcome: rater1, rater2, rater3
#>
#> Assumptions
#> * Independence of observations: assumed: Subjects are assumed to be measured independently of one another
#>
#> Recommended test
#> Intraclass correlation coefficient
#>
#> Result
#> H0: ICC(2,1) = 0 (no reliability beyond chance).
#> statistic = 1.68, df = 29.00, p = 0.046, 95% CI [-0.03, 0.44]
#>
#> Effect size
#> ICC(2,1): 0.19, poor
#>
#> ICC comparison
#>
#>
#> method estimate conf.low conf.high statistic p.value
#> ICC(1,1) one-way random 0.20 -0.02 0.44 1.73 0.037
#> ICC(2,1) two-way random, absolute agreement 0.19 -0.03 0.44 1.68 0.046
#> ICC(3,1) two-way fixed, consistency 0.19 -0.03 0.43 1.68 0.046
#>
#> Report
#> The intraclass correlation workflow for rater1, rater2, rater3 showed a statistically significant result
```

`test_icc()` reports ICC(2,1) (two-way random effects, absolute agreement, single measurement) as the primary result, following the reliability-study recommendation of Koo & Li (2016), alongside ICC(1,1) and ICC(3,1) for comparison in `alternative_tests$icc_table`.

Outliers

```
test_outliers(c(sbp_3m, ldl, crp), data = cardio)
#> Warning: `outliers` is a screening workflow, not a single hypothesis test.
```

```
#> Outlier Screening
#>
#> Outcome: sbp_3m, ldl, crp
#>
#> Assumptions
#> * Numeric variable: acceptable: IQR outlier detection is univariate and does not require normality.
#> * Skewness sensitivity: warning: Interpret IQR outliers with care when the distribution is strongly .
#>
#> Recommended test
#> IQR outlier detection
#>
#> Result
#> flagged rows = 11
#>
#> Effect size
#> Outlier count: 11.00
#>
#> Report
#> The outlier workflow flagged 11 rows for review.
```

Reporting and plotting

Every workflow returns a `testflow` object. Use `report(x)`, `plot(x)`, and `as_tibble(x)`. See `effect-size-formulas.Rmd` for the exact formulas used by the reported effect-size estimates.