testflow: complete pedagogical reference

```
library(testflow)
cardio <- make_cardio_data()
```

Introduction

This vignette is the complete pedagogical reference for `testflow`: every exported workflow function, organized by study design, alongside the exact mathematical formula each one evaluates. It supersedes and consolidates the previous `effect-size-formulas` and `statistical-test-workflows` vignettes, and folds in the sample-size planning module so that every formula in the package - hypothesis testing, effect sizes, and sample-size planning - lives in one place.

Every `test_*()` function returns a `testflow` object: a list containing the cleaned data, descriptive statistics, assumption checks, the recommended test with its rationale, the primary test result (with its null hypothesis), alternative test results for comparison, an effect size, an optional `ggplot`, the original call, and a plain-language interpretation. Five methods work on every object:

```
x <- test_two_groups(sbp_3m ~ sex, data = cardio)
x                # print.testflow(): formatted console report
#> 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-Wilk; statistic=0.99; p=0.913)
#> * Normality: sbp_3m (male): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.98; p=0.233)
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable.
↪  (method=Levene test; statistic=1.57)
#> * 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 Student independent t-test, statistic = -1.91, df = 178.00,
↪  p = 0.058. The 95% confidence interval was [-11.22, 0.18]. The effect size was small
↪  (Cohen's d = -0.29). H0: the population mean or location of sbp_3m is equal across
↪  levels of sex.
```

```r
summary(x)       # summary.testflow(): compact summary list
#> Two Independent Groups (summary)
#>
#> Workflow: two_groups
#> Outcome: sbp_3m
#> Group: sex
#> Recommended test: Student independent t-test
#> H0: H0: the population mean or location of sbp_3m is equal across levels of sex.
#>
#> Result
#> Statistic: -1.91
#> df: 178.00
#> p: 0.058
#> 95% CI: [-11.22, 0.18]
#> Effect size: Cohen's d = -0.29, small
#> Decision: Not statistically significant at the chosen alpha level.
#>
#> Report
#> The two independent groups workflow for sbp_3m did not show a statistically
↪  significant result using Student independent t-test, statistic = -1.91, df = 178.00,
↪  p = 0.058. The 95% confidence interval was [-11.22, 0.18]. The effect size was small
↪  (Cohen's d = -0.29). H0: the population mean or location of sbp_3m is equal across
↪  levels of sex.
report(x)        # report.testflow(): report-ready sentence
#> [1] "The two independent groups workflow for sbp_3m did not show a statistically
↪  significant result using Student independent t-test, statistic = -1.91, df = 178.00,
↪  p = 0.058. The 95% confidence interval was [-11.22, 0.18]. The effect size was small
↪  (Cohen's d = -0.29). H0: the population mean or location of sbp_3m is equal across
↪  levels of sex."
as_tibble(x)     # as_tibble.testflow(): one-row tidy summary
#> # A tibble: 1 x 15
#>   workflow design outcome group recommended_test null_hypothesis statistic    df
#>   <chr>    <chr>  <chr>   <chr> <chr>            <chr>               <dbl> <dbl>
#> 1 two_gro~ two i~ sbp_3m  sex   Student indepen~ H0: the popula~     -1.91   178
#> # i 7 more variables: p <dbl>, conf.low <dbl>, conf.high <dbl>,
#> #   effect_size <dbl>, effect_size_name <chr>, effect_size_magnitude <chr>,
#> #   decision <chr>
```

`plot(x)` produces the workflow's `ggplot`; every workflow's own plot is shown alongside its formula in the sections below.

`report_test(x)` is a plain-function alias for `report(x)`, provided for callers who prefer not to rely on S3 dispatch. `make_cardio_data(n, seed)` (used throughout this vignette) simulates the small teaching dataset used in every non-survival, non-diagnostic, non-agreement example below - it is a data generator, not a statistical workflow, and carries no formula of its own.

Notation used throughout: n is a sample size, $\bar{x}$ and s_x are a sample mean and standard deviation, z_p and $t_{df,p}$ are the p-th quantiles of the standard normal and t distributions, and χ^2_{df}, F_{df_1,df_2} name a chi-square and F distribution with the given degrees of freedom.

How test selection works

Most `test_*()` functions run assumption checks first and use them to pick among two or three candidate tests (the *recommended* test), while still computing and reporting every candidate under `alternative_tests` so you can compare. The decision rules (`R/recommend.R`) are simple and deliberately transparent:

Workflow	Normal? & equal variance?	Normal only	Neither
One sample	One-sample t-test	-	Wilcoxon signed-rank
Two groups	Student t-test	Welch t-test	Wilcoxon rank-sum
Paired	Paired t-test	-	Wilcoxon signed-rank
3+ groups	One-way ANOVA	Welch ANOVA	Kruskal-Wallis

Normality is checked with Shapiro-Wilk (per group, where applicable); variance homogeneity with Levene's test. `test_categorical()` instead compares chi-square vs. Fisher's exact test based on expected cell counts (`fisher_threshold`, default 5); `test_correlation(method = "auto")` compares Pearson vs. Spearman based on normality and outlier screens; the numeric repeated-measures workflow compares repeated-measures ANOVA vs. Friedman based on within-time normality.

Descriptive statistics

`sumtab()` and the internal `descriptives_numeric()`/`descriptives_categorical()` helpers (populating every `testflow` object's `$descriptives` field) report, for numeric variables:

$$n, \quad \bar{x} = \frac{1}{n}\sum_i x_i, \quad s_x = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}, \quad \text{median}, \quad IQR = Q_3 - Q_1, \quad \text{min}, \ \text{max}$$

and, for categorical variables, the count and percentage of each level:

```
sumtab(~ age + sbp_3m + ldl | treatment, data = cardio)
#>  variable level                          Overall (n = 180)                      usual
↪  care (n = 55)                        lifestyle (n = 71)
↪  medication (n = 54)
#>       age  <NA>     57.9 (10.7); 58.0 [51.0, 64.2]; n=180     56.3 (10.0); 56.0 [51.0,
↪  61.0]; n=55     59.4 (10.4); 60.0 [52.0, 67.0]; n=71     57.6 (11.7); 57.5 [50.0,
↪  64.0]; n=54
#>    sbp_3m  <NA> 139.3 (19.5); 139.8 [127.5, 151.0]; n=180 144.1 (19.3); 146.0 [132.8,
↪  152.7]; n=55 140.1 (18.2); 139.4 [132.4, 150.9]; n=71 133.5 (20.1); 135.4 [119.1,
↪  148.2]; n=54
#>       ldl  <NA>          3.3 (0.8); 3.3 [2.8, 3.8]; n=180          3.2 (0.9); 3.2
↪  [2.5, 3.9]; n=55          3.3 (0.7); 3.4 [2.9, 3.8]; n=71          3.1 (0.7); 3.2
↪  [2.8, 3.6]; n=54
```

Assumption checks

`testflow` runs a small set of standardized assumption checks and reports them in every object's `$assumptions` field, each with a `status` (`acceptable`/`not acceptable`/`warning`/`not applicable`/`assumed`/`not checked`) driving the recommended-test logic above.

Check	Method	Statistic
Normality	Shapiro-Wilk (`n <= 5000`)	W
Symmetry (one-sample/paired)	Cabilio-Masaro	$z = \sqrt{n}(\bar{x} - \text{median})/(s_x\sqrt{\pi/2 - 1})$
Variance homogeneity (2 groups)	`var.test()` (F-test)	$F = s_1^2/s_2^2$
Variance homogeneity (3+ groups)	Levene's test	F on absolute deviations from group medians
Variance homogeneity (3+ groups)	Bartlett's test	χ^2 on group variances

Check	Method	Statistic
Sphericity (3+ repeated measures)	Mauchly's test	W
Linearity (correlation)	Quadratic-term F-test (RESET, power 2)	F comparing `y ~ x` to `y ~ x + I(x^2)`
Monotonicity (correlation)	Spearman correlation of x, y	ρ
Expected cell counts (categorical)	Pearson chi-square expected counts	$\min(E_{ij}) \geq$ threshold (default 5)
Multicollinearity (regression)	Variance inflation factor	$VIF_j = 1/(1 - R_j^2)$, flagged at ≥ 5
Influential points (logistic regression)	Cook's distance	flagged at $> 4/n$
Proportional hazards (Cox)	Schoenfeld residual test (Grambsch-Therneau)	χ^2
Outliers (screen)	IQR rule	$x < Q_1 - 1.5\,IQR$ or $x > Q_3 + 1.5\,IQR$

Independence of observations is *assumed* from the study design in every workflow (`testflow` has no way to test it from the data alone) and is always listed explicitly rather than silently taken for granted.

One numerical variable

`test_one_sample()` compares a numeric sample against a fixed reference value μ_0. The recommended test follows normality of the sample:

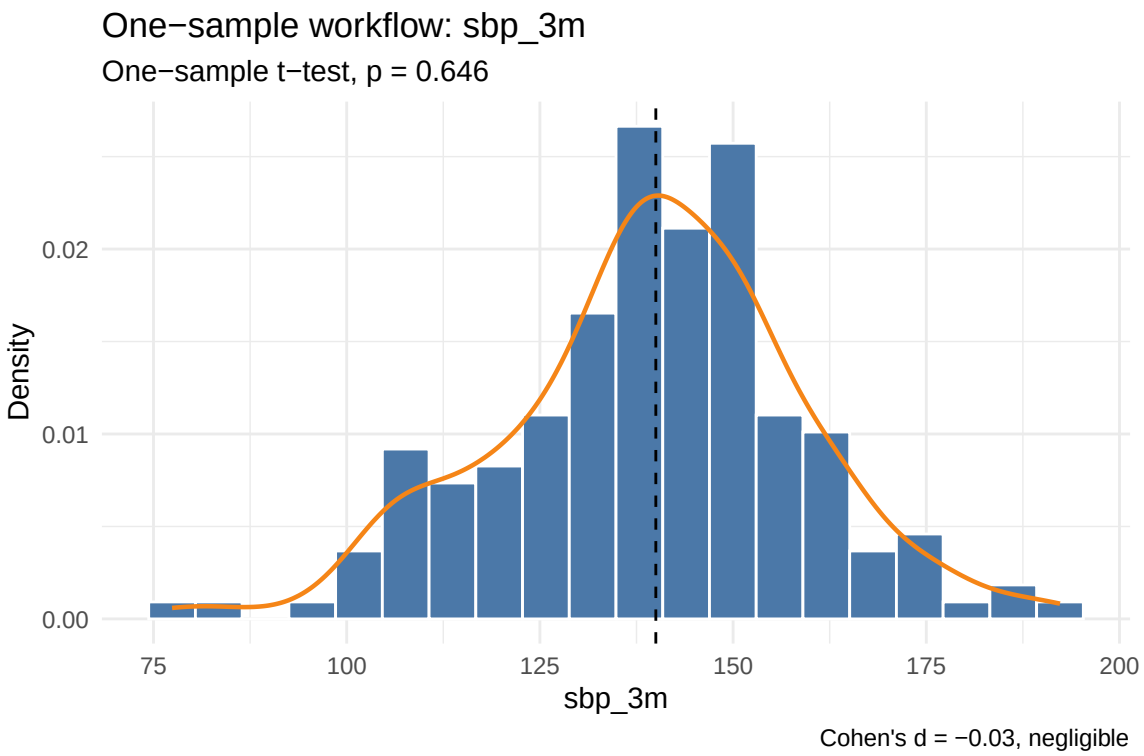
$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} \sim t_{n-1} \qquad \text{(One-sample t-test, Gosset 1908)}$$

with the Wilcoxon signed-rank test as the non-parametric alternative, and a sign test always reported alongside for reference. The reported effect size is Cohen's d:

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

```
x_one_sample <- test_one_sample(cardio, sbp_3m, mu = 140)
x_one_sample
#> 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; statistic=0.99; p=0.308)
#> * Symmetry of deviations: acceptable: The deviations from the reference value look
↪  approximately symmetric. (method=Cabilio-Masaro test; statistic=-0.38; p=0.703)
#>
#> 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 One-sample t-test, statistic = -0.46, df = 179.00, p = 0.646. The 95%
↪  confidence interval was [136.47, 142.20]. The effect size was negligible (Cohen's d =
↪  -0.03). H0: the population mean or location of sbp_3m equals the reference value.
plot(x_one_sample)
```



Two independent groups

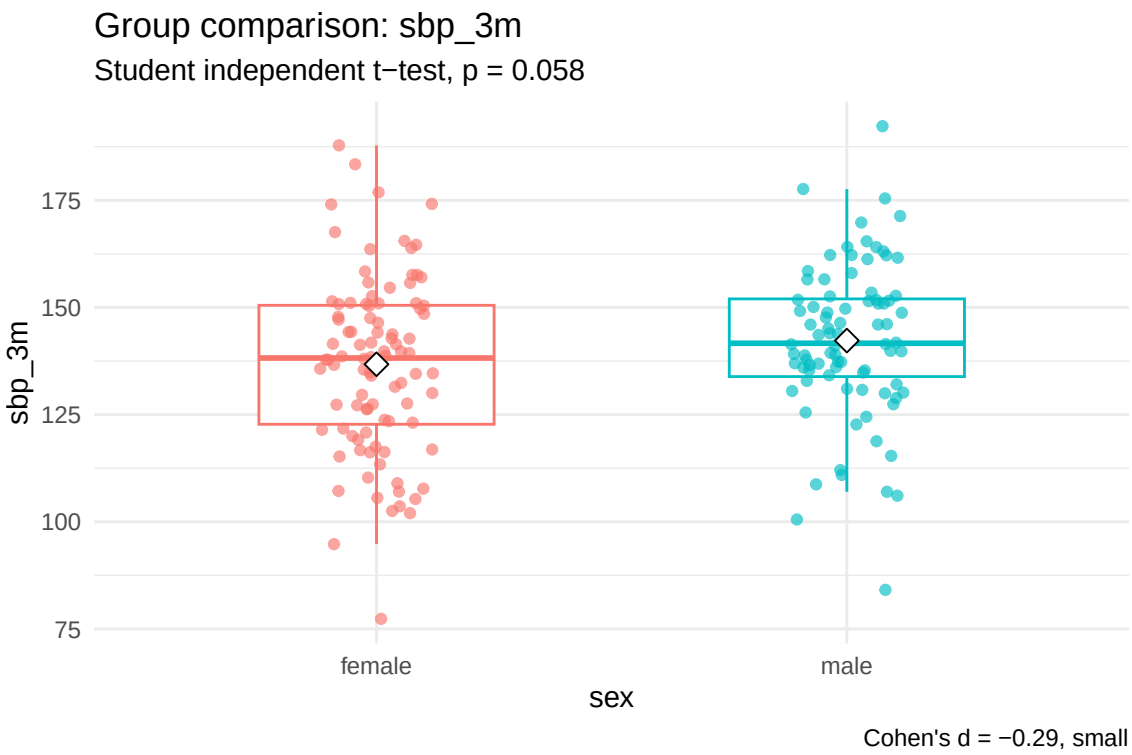
`test_two_groups()` compares a numeric outcome between two independent groups, selecting among the Student t-test, Welch t-test, and Wilcoxon rank-sum test per the table above:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}, \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \qquad \text{(Student/Gosset 1908)}$$

Welch's t-test (1947) instead uses $s_1^2/n_1 + s_2^2/n_2$ as the variance estimate with a Satterthwaite-approximated denominator degrees of freedom - appropriate when variances are unequal. The Wilcoxon rank-sum (Mann-Whitney) test compares rank sums rather than means. Effect size is Cohen's d (t-test branches) or the rank-biserial correlation $r_{rb} = 1 - 2W/(n_1 n_2)$ (Wilcoxon branch), both with the sign following "first group minus second group."

```
x_two_groups <- test_two_groups(sbp_3m ~ sex, data = cardio)
x_two_groups
#> 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-Wilk; statistic=0.99; p=0.913)
#> * Normality: sbp_3m (male): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.98; p=0.233)
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable.
↪  (method=Levene test; statistic=1.57)
#> * 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 Student independent t-test, statistic = -1.91, df = 178.00,
↪  p = 0.058. The 95% confidence interval was [-11.22, 0.18]. The effect size was small
↪  (Cohen's d = -0.29). H0: the population mean or location of sbp_3m is equal across
↪  levels of sex.
plot(x_two_groups)
```



Paired measurements

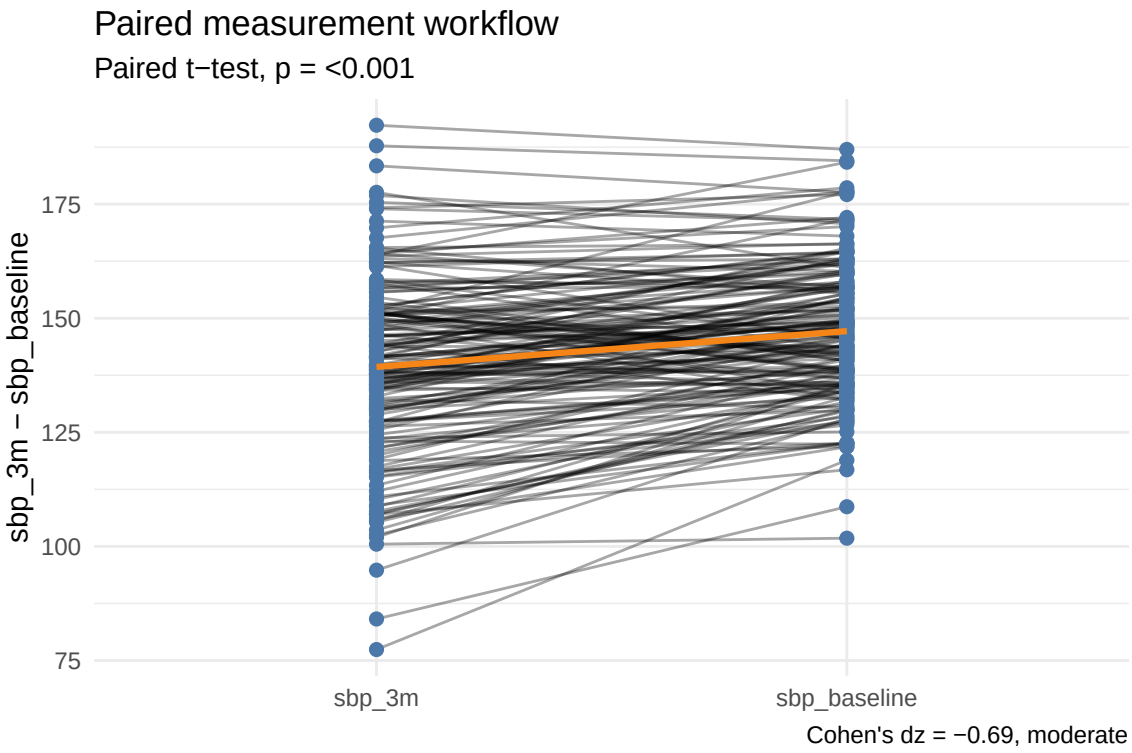
`test_paired()` compares two measurements on the same subjects via their paired differences $d_i = \text{after}_i - \text{before}_i$:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1} \qquad \text{(Paired t-test)}$$

with the Wilcoxon signed-rank test on the differences as the non-parametric alternative (selected when the differences fail the normality check), and a sign test reported for reference. Effect size is Cohen's $d_z = \bar{d}/s_d$.

```
x_paired <- test_paired(sbp_3m ~ sbp_baseline, data = cardio)
x_paired
#> Paired Measurements
#>
#> Outcome: sbp_3m - sbp_baseline
#>
#> Assumptions
#> * Independence of observations: assumed: Paired observations from the same subjects
↪  are assumed by design.
#> * Normality: diff: acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.99; p=0.557)
#> * Symmetry of paired differences: acceptable: The deviations from the reference value
↪  look approximately symmetric. (method=Cabilio-Masaro test; statistic=0.47; p=0.635)
#> * 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 using Paired t-test, statistic = -9.20, df = 179.00, p = <0.001.
↪  The 95% confidence interval was [-9.53, -6.16]. The effect size was moderate (Cohen's
↪  dz = -0.69). H0: the mean or median paired difference (sbp_3m - sbp_baseline) equals
↪  0.
plot(x_paired)
```



More than two groups

`test_groups()` extends the two-group comparison to $k \geq 3$ groups:

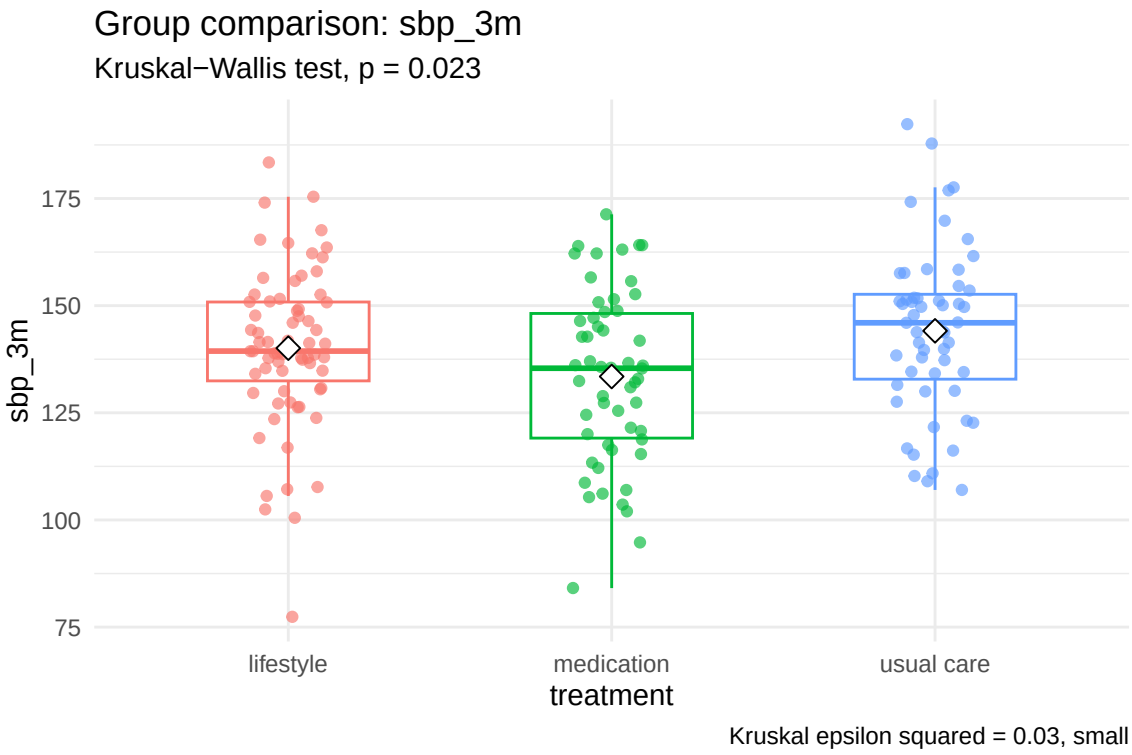
$$F = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)} \sim F_{k-1,\,n-k} \qquad \text{(One-way ANOVA, Fisher 1925)}$$

Welch's ANOVA relaxes the equal-variance assumption (analogous to Welch's t-test); the Kruskal-Wallis test is the rank-based, distribution-free alternative:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1) \sim \chi^2_{k-1} \qquad \text{(Kruskal-Wallis, 1952)}$$

Effect size is eta squared $\eta^2 = SS_{between}/SS_{total}$ for the ANOVA branches, or Kruskal's epsilon squared $\epsilon^2 = (H - k + 1)/(n - k)$ for the Kruskal-Wallis branch.

```
x_groups <- test_groups(sbp_3m ~ treatment, data = cardio)
x_groups
#> 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; statistic=0.96; p=0.030)
#> * Normality: sbp_3m (medication): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.98; p=0.647)
#> * Normality: sbp_3m (usual care): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.98; p=0.349)
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable.
↪  (method=Levene test; statistic=1.20)
#> * Bartlett test: acceptable: Variance homogeneity looks reasonable. (method=Bartlett
↪  test; statistic=0.66)
#> * 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 using Kruskal-Wallis test, statistic = 7.58, df = 2.00, p = 0.023.
↪  The effect size was small (Kruskal epsilon squared = 0.03). H0: the population mean
↪  or location of sbp_3m is equal across levels of treatment.
plot(x_groups)
```



Post-hoc comparisons

When `posthoc = TRUE` (the default), pairwise comparisons follow the selected omnibus test: Tukey's HSD (1949) after one-way ANOVA, BH-adjusted pairwise Welch t-tests after Welch ANOVA, and BH-adjusted pairwise Wilcoxon tests (Dunn-style multiple comparisons, 1964) after Kruskal-Wallis.

Factorial designs

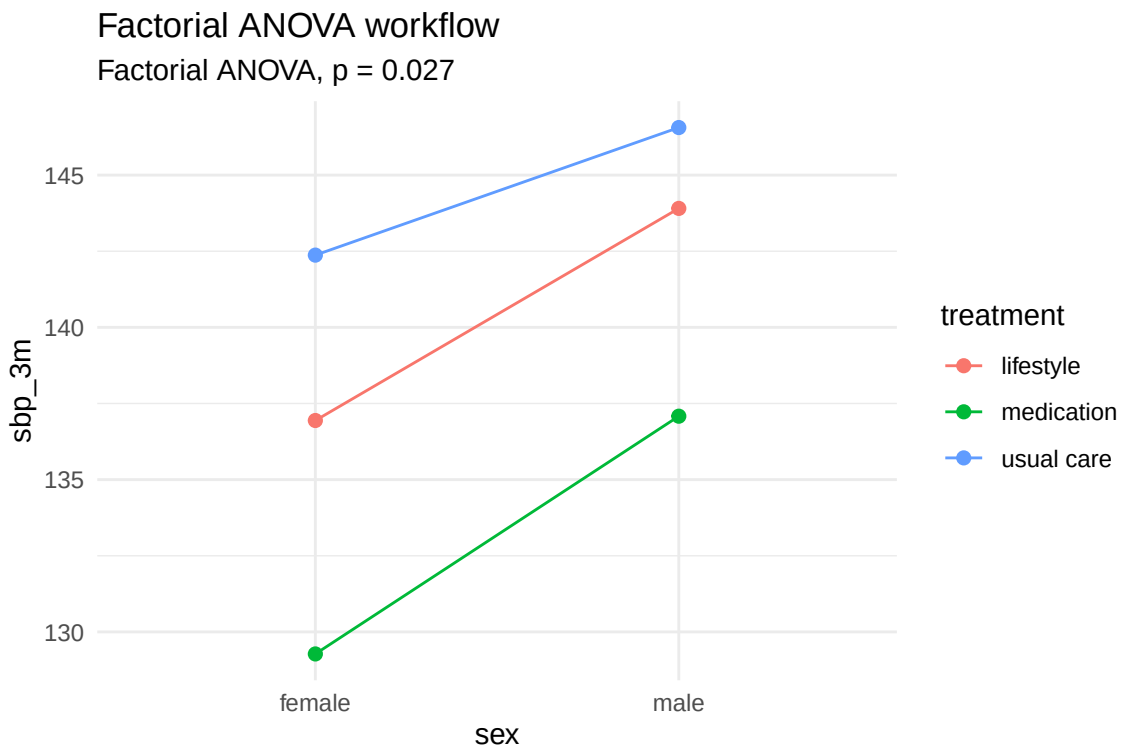
`test_factorial()` fits a two- (or more-) way ANOVA with `outcome ~ f1 * f2` and reports sums of squares of the chosen `type`:

- **Type I** (`stats::anova()`): sequential SS, order-dependent for unbalanced designs.
- **Type II/Type III** (`car::Anova()`): each term's SS adjusted for every other term (Type II) or additionally for the intercept under sum-to-zero contrasts (Type III, `testflow` sets `contr.sum` automatically since Type III under the default treatment contrasts tests a different, usually meaningless, hypothesis).

Effect size is eta squared per term, computed from whichever SS table (Type I/II/III) was actually used for the reported test - not always recomputed as Type I - so the effect size matches the p-value. Residual normality (Shapiro-Wilk) and variance homogeneity (Levene, on the first factor) are checked as assumptions.

```
x_factorial <- test_factorial(sbp_3m ~ sex * treatment, data = cardio)
x_factorial
#> 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; statistic=0.99; p=0.560)
```

```
#> * Variance homogeneity: acceptable: Variance homogeneity looks reasonable.
↪  (method=Levene test; statistic=1.57; p=0.211; Df1=1; Df2=178)
#> * Balanced design: not required: Cell sizes are unbalanced; Type II sums of squares
↪  are used so the terms are not order-dependent.
#>
#> 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 ANOVA, statistic = 5.01, df = 1.00, p = 0.027. The effect size was
↪  small (eta squared = 0.03). H0: the population mean or location of sbp_3m is equal
↪  across levels of sex, treatment.
plot(x_factorial)
```



Repeated measurements (numeric)

`test_repeated()` (wide data) and `test_repeated_long()` (long data) compare 3 or more repeated numeric measurements on the same subjects, choosing between repeated-measures ANOVA and the Friedman test based on within-time normality:

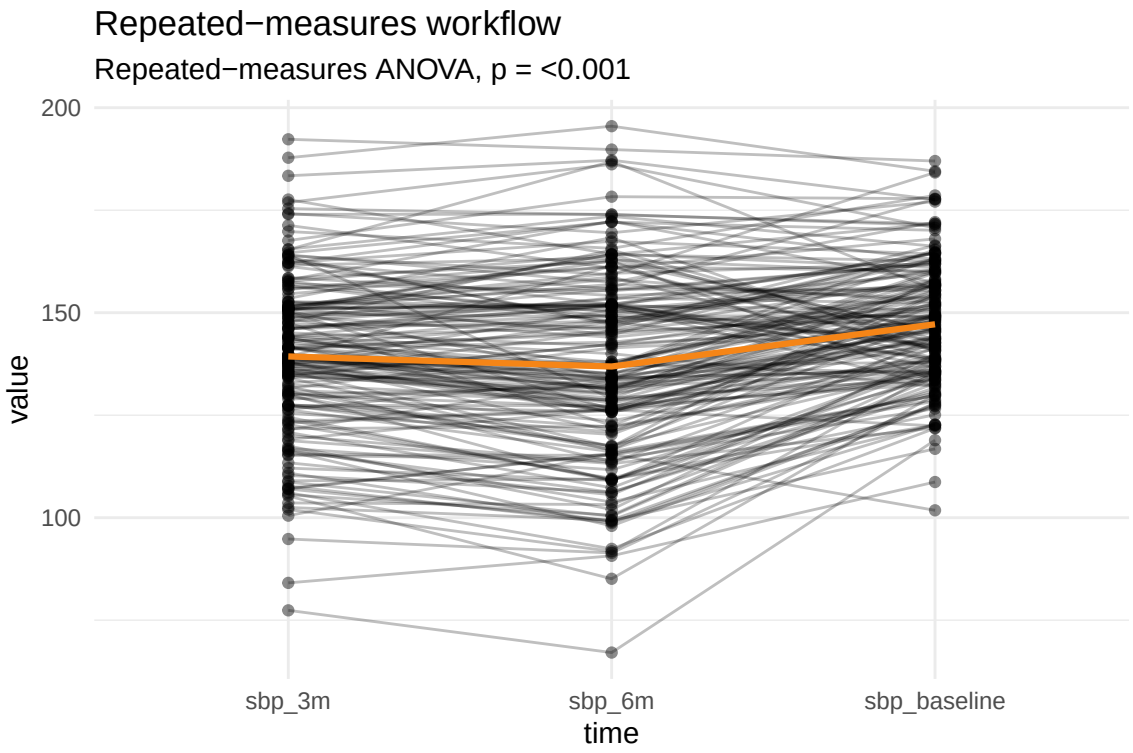
$$F = \frac{MS_{time}}{MS_{time \times subject}} \sim F_{k-1,\,(k-1)(n-1)} \qquad \text{(Repeated-measures ANOVA, via 'Error(id/time)')}$$

$$\chi^2_F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \sim \chi^2_{k-1} \qquad \text{(Friedman, 1937)}$$

Sphericity (Mauchly's test) is checked for the ANOVA branch. Effect size is $\eta^2_{time} = SS_{time}/SS_{total}$ (ANOVA branch) or Kendall's $W = \chi^2_F/(n(k-1))$ (Friedman branch). Post-hoc comparisons are BH-adjusted pairwise paired t-tests (ANOVA branch) or pairwise paired Wilcoxon tests (Friedman branch).

```
x_repeated <- test_repeated(cardio, c(sbp_baseline, sbp_3m, sbp_6m), id = id)
x_repeated
#> 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 design.
#> * Normality: sbp_3m: acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.99; p=0.308)
#> * Normality: sbp_6m: acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=1.00; p=0.842)
#> * Normality: sbp_baseline: acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.99; p=0.732)
#> * Sphericity: warning: Sphericity may be violated; interpret the repeated-measures
↪  ANOVA with caution, or prefer the Friedman test. (method=Mauchly's test;
↪  statistic=0.64; p=<0.001)
#>
#> 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 significant result using Repeated-measures ANOVA, statistic = 68.14, df
↪  = 2.00, p = <0.001. The effect size was small (eta squared = 0.05). H0: the
↪  population mean or location of sbp_baseline, sbp_3m, sbp_6m is equal across levels of
↪  time.
plot(x_repeated)
```



`test_repeated_long()` runs the identical workflow (same formulas, same recommendation logic, same plot) starting from long-format data instead of wide: pass the outcome, the within-subject time/condition column, and the subject id column directly, rather than a set of wide columns to pivot.

```
long_sbp <- data.frame(
  id = rep(cardio$id, 3),
  time = rep(c("sbp_baseline", "sbp_3m", "sbp_6m"), each = nrow(cardio)),
  sbp = c(cardio$sbp_baseline, cardio$sbp_3m, cardio$sbp_6m)
)
test_repeated_long(long_sbp, outcome = sbp, within = time, id = id)
#> Repeated Measures
#>
#> Outcome: sbp
#> Group: time
#>
#> Assumptions
#> * Independence of observations: assumed: Repeated measurements from the same subjects
↪  are assumed by design.
#> * Normality: sbp (sbp_3m): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.99; p=0.308)
#> * Normality: sbp (sbp_6m): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=1.00; p=0.842)
```

```
#> * Normality: sbp (sbp_baseline): acceptable: Approximate normality looks reasonable.
↪  (method=Shapiro-Wilk; statistic=0.99; p=0.732)
#> * Sphericity: warning: Sphericity may be violated; interpret the repeated-measures
↪  ANOVA with caution, or prefer the Friedman test. (method=Mauchly's test;
↪  statistic=0.64; p=<0.001)
#>
#> Recommended test
#> Repeated-measures ANOVA
#>
#> Result
#> H0: the population mean or location of sbp 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 showed a statistically significant
↪  result using Repeated-measures ANOVA, statistic = 68.14, df = 2.00, p = <0.001. The
↪  effect size was small (eta squared = 0.05). H0: the population mean or location of
↪  sbp is equal across levels of time.
```

Categorical outcomes

Association between two categorical variables

`test_categorical()` tests independence in a contingency table, selecting chi-square or Fisher's exact test based on expected cell counts:

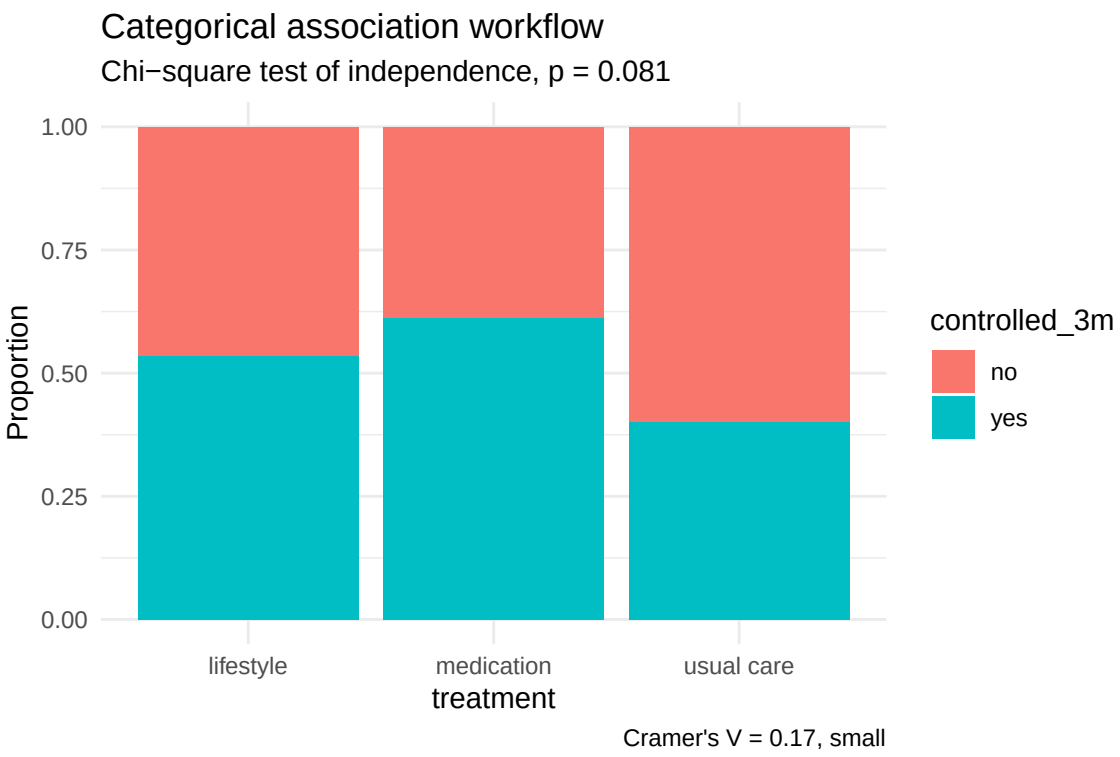
$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \qquad \text{(Pearson 1900)}$$

Fisher's exact test (1922) is used instead whenever any expected count falls below `fisher_threshold` (default 5). Effect size is Cramer's V:

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

```
x_categorical <- test_categorical(treatment ~ controlled_3m, data = cardio)
x_categorical
#> 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-square approximation; Min expected = 26.1; threshold = 5)
#>
#> 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 using Chi-square test of independence, statistic = 5.02, df =
↪  2.00, p = 0.081. The effect size was small (Cramer's V = 0.17). H0: treatment and
↪  controlled_3m are independent.
plot(x_categorical)
```

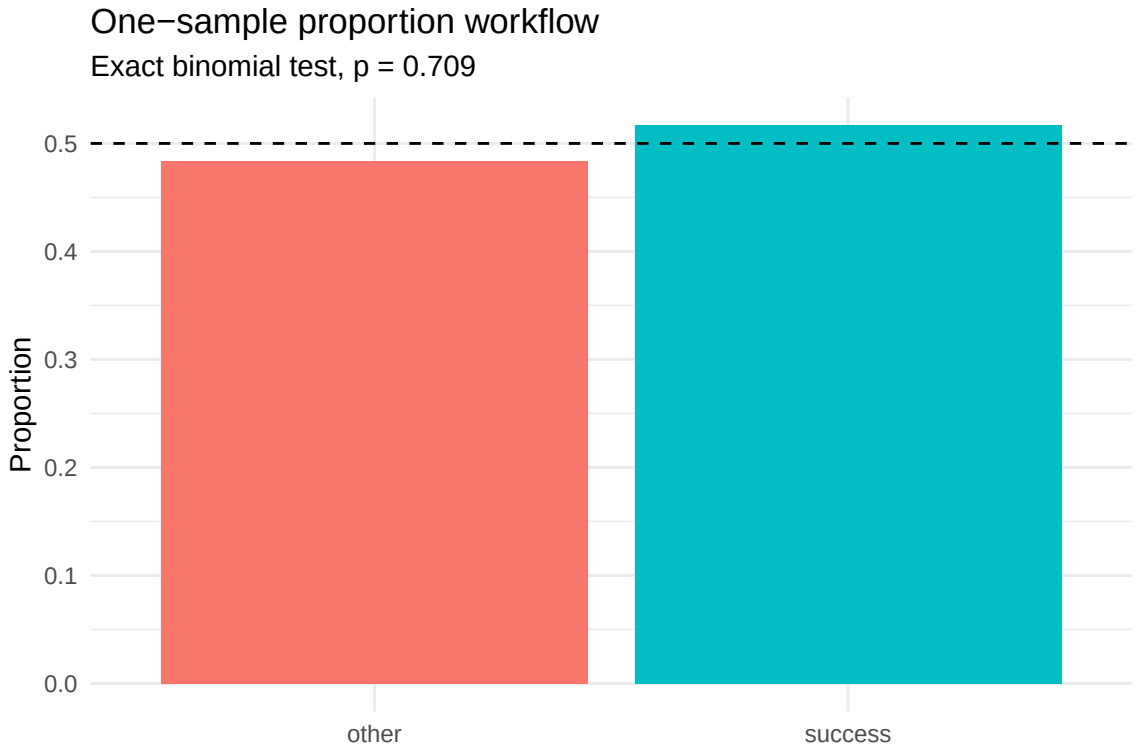


One-sample proportion

`test_proportion()` tests a single proportion against a reference value p_0, using the exact binomial test (Clopper-Pearson-based) when expected successes or failures fall below 5, or the standard proportion test otherwise:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \qquad \text{(one-sample proportion test, normal approximation)}$$

```
x_proportion <- test_proportion(cardio, controlled_3m, success = "yes", p = 0.5)
x_proportion
#> One-Sample Proportion
#>
#> Outcome: controlled_3m
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Fixed number of trials: assumed: The workflow treats the observed sample size as
↪  fixed.
#> * Expected successes and failures: acceptable: Approximate one-sample proportion test
↪  is reasonable. (expected_success = 90; expected_failure = 90)
#>
#> Recommended test
#> One-sample proportion test
#>
#> Result
#> H0: the true proportion of controlled_3m successes equals 0.5.
#> statistic = 0.20, df = 1.00, p = 0.655, 95% CI [0.44, 0.59]
#>
#> Effect size
#> Observed proportion: 0.52
#>
#> Report
#> The one categorical proportion workflow for controlled_3m did not show a statistically
↪  significant result using One-sample proportion test, statistic = 0.20, df = 1.00, p =
↪  0.655. The 95% confidence interval was [0.44, 0.59]. The effect size (Observed
↪  proportion) was 0.52. H0: the true proportion of controlled_3m successes equals 0.5.
plot(x_proportion)
```



Multinomial goodness of fit

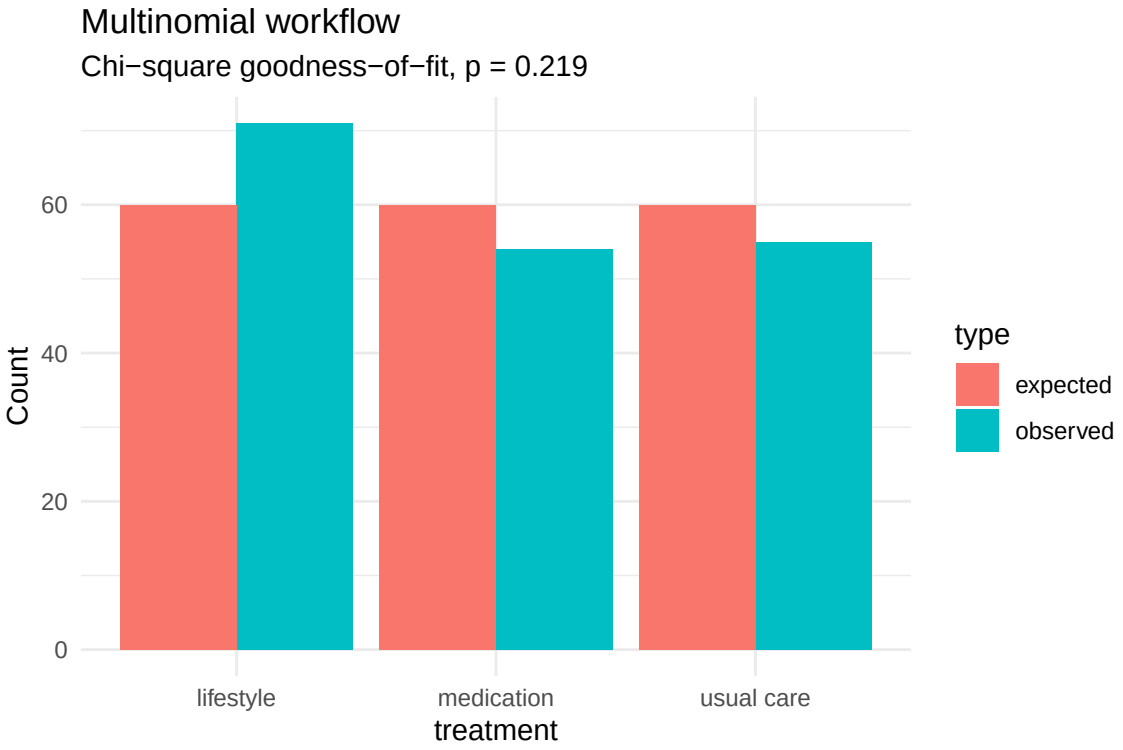
`test_multinomial()` compares observed category counts to expected probabilities $p_1, \ldots, p_k$ (equal by default):

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

BH-adjusted per-category binomial tests are reported alongside as post-hoc comparisons.

```
x_multinomial <- test_multinomial(cardio, treatment)
#> Warning: No expected probabilities supplied; using equal probabilities.
x_multinomial
#> Multinomial Goodness of Fit
#>
#> Outcome: treatment
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Expected category counts: acceptable: Chi-square approximation is reasonable. (min
↪  expected = 60)
#>
#> Recommended test
#> Chi-square goodness-of-fit
#>
#> Result
#> H0: the distribution of treatment follows the expected probabilities.
#> statistic = 3.03, df = 2.00, p = 0.219
#>
```

```
#> Effect size
#> Chi-square goodness-of-fit: 3.03
#>
#> Pairwise category checks
#>       level observed expected     p p.adj p.adjust.method
#>   lifestyle    71.00    60.00 0.096 0.289              BH
#>  medication    54.00    60.00 0.385 0.477              BH
#>  usual care    55.00    60.00 0.477 0.477              BH
#>
#> Report
#> The one multinomial categorical variable workflow for treatment did not show a
↪  statistically significant result using Chi-square goodness-of-fit, statistic = 3.03,
↪  df = 2.00, p = 0.219. The effect size (Chi-square goodness-of-fit) was 3.03. H0: the
↪  distribution of treatment follows the expected probabilities.
plot(x_multinomial)
```



Paired categorical measurements

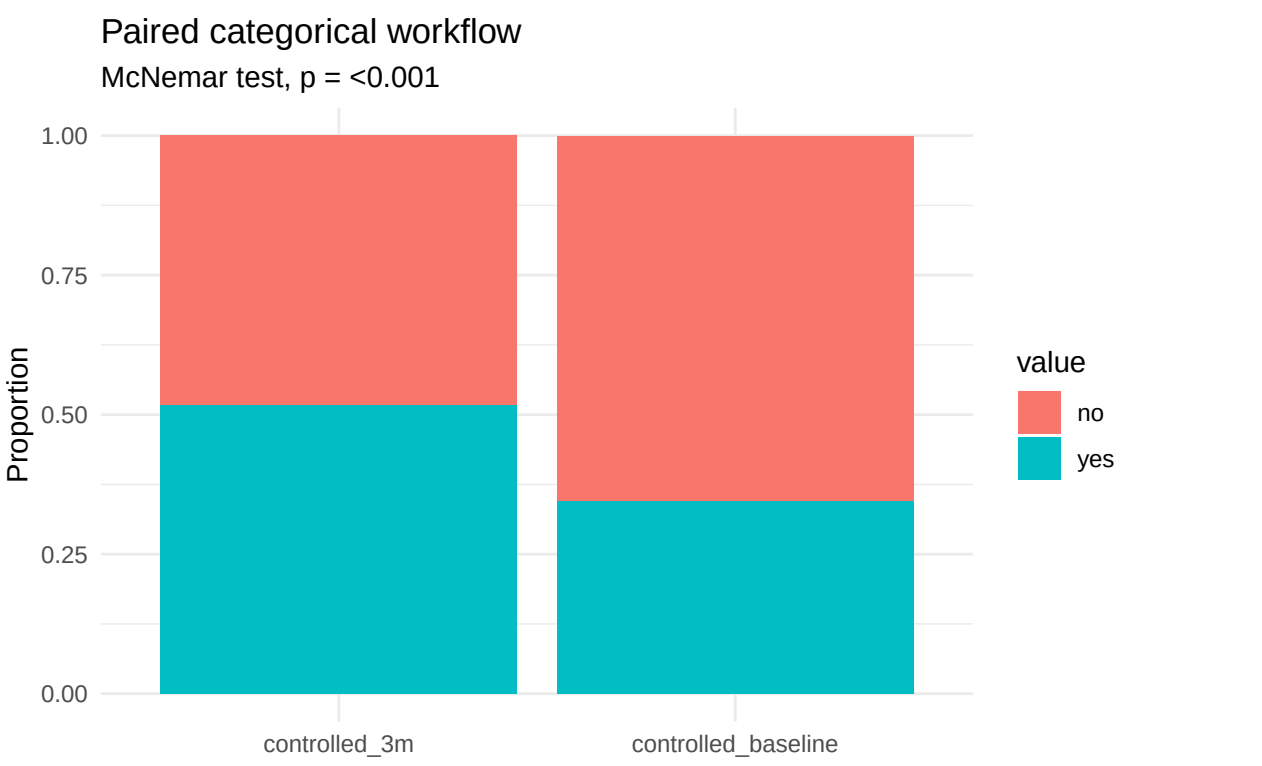
`test_paired_categorical()` compares a binary categorical outcome measured twice on the same subjects via McNemar's test on the discordant cells b (before-only) and c (after-only) of the 2x2 table:

$$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi^2_1 \qquad \text{(McNemar, 1947)}$$

Effect size is the number of discordant pairs $b + c$.

```
x_paired_cat <- test_paired_categorical(cardio, controlled_baseline, controlled_3m)
x_paired_cat
#> Paired Categorical Measurements
#>
```

```
#> Outcome: controlled_baseline -> controlled_3m
#>
#> Assumptions
#> * Paired binary measurements: assumed: Same subjects should be measured twice.
#> * Discordant pairs: acceptable: Discordant pairs are adequate for the McNemar
↪  approximation. (discordant = 180)
#>
#> Recommended test
#> McNemar test
#>
#> Result
#> statistic = 18.37, df = 1.00, p = <0.001
#>
#> Effect size
#> Discordant pairs: 49.00
#>
#> Discordant pairs
#>  before_only after_only
#>            9         40
#>
#> Report
#> The paired categorical measurements workflow for controlled_baseline -> controlled_3m
↪  showed a statistically significant result using McNemar test, statistic = 18.37, df =
↪  1.00, p = <0.001. The effect size (Discordant pairs) was 49.00.
plot(x_paired_cat)
```



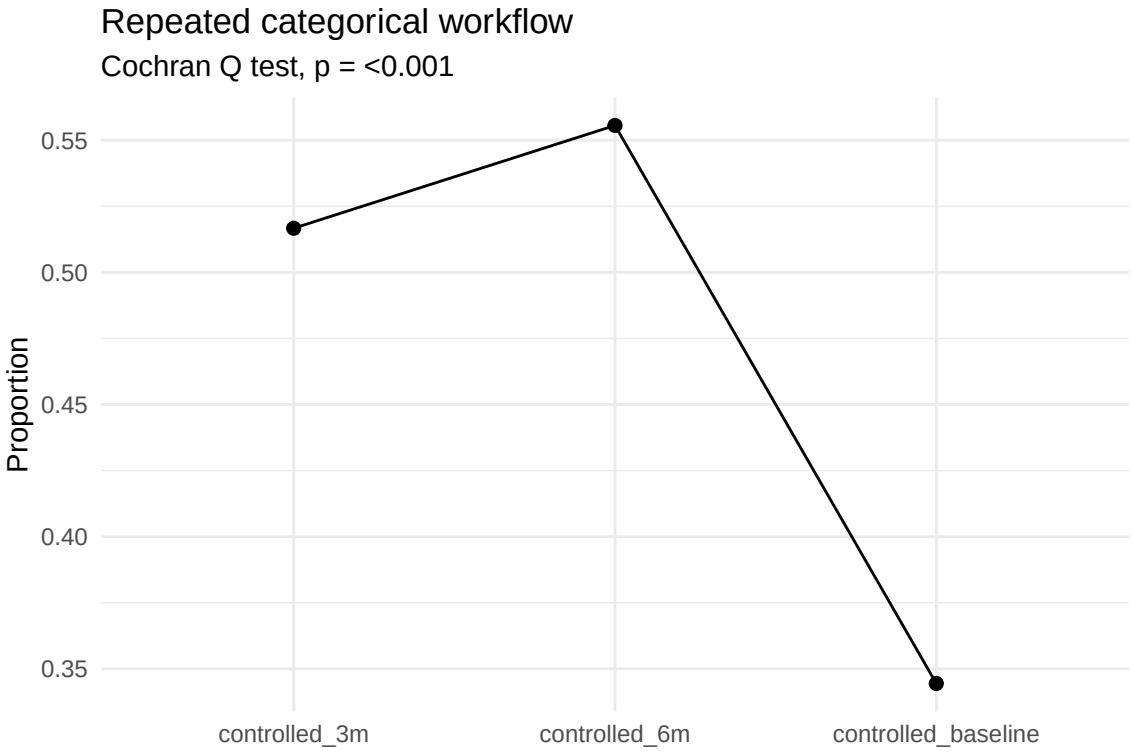
Repeated categorical measurements (3+ time points)

`test_repeated_categorical()` extends McNemar's test to $k \geq 3$ repeated binary measurements via Cochran's Q test:

$$Q = \frac{(k-1)\left[k\sum_j C_j^2 - \left(\sum_j C_j\right)^2\right]}{k\sum_i R_i - \sum_i R_i^2} \sim \chi^2_{k-1} \qquad \text{(Cochran, 1950)}$$

where C_j is the column (time) total and R_i the row (subject) total. Effect size is the Cochran-Q analogue of Kendall's W, $W = Q/(n(k-1))$; pairwise McNemar tests (BH-adjusted) are the post-hoc comparisons.

```
x_repeated_cat <- test_repeated_categorical(cardio, c(controlled_baseline, controlled_3m,
↪  controlled_6m))
x_repeated_cat
#> 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 analysis.
#>
#> 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 showed a statistically significant result using Cochran Q test,
↪  statistic = 39.58, df = 2.00, p = <0.001. The effect size was small (Cochran Q
↪  Kendall's W = 0.11). H0: the success proportions are equal across repeated
↪  categorical measures.
plot(x_repeated_cat)
```



Correlation

`test_correlation()` tests association between two numeric variables, with `method = "auto"` choosing Pearson when both variables pass the normality and outlier screens, and Spearman otherwise:

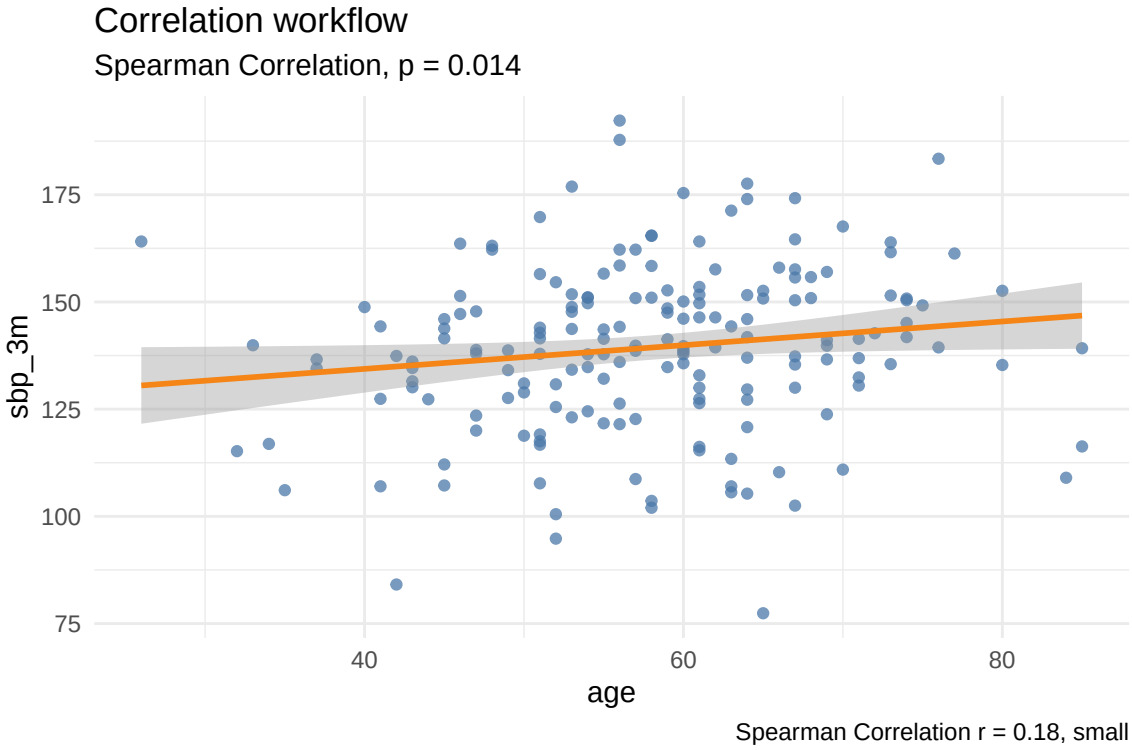
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad \text{(Pearson, 1895)}$$

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}, \quad d_i = \text{rank}(x_i) - \text{rank}(y_i) \qquad \text{(Spearman, 1904)}$$

Kendall's τ (1938), based on the number of concordant vs. discordant pairs, is always computed and reported in `alternative_tests`. Each of the three is a candidate effect size, reported as the estimate of the selected method. Linearity (for Pearson) and monotonicity (for Spearman/Kendall) are checked as assumptions.

```
x_correlation <- test_correlation(sbp_3m ~ age, data = cardio)
x_correlation
#> Correlation
#>
#> Outcome: sbp_3m
#> Group: age
#>
#> Assumptions
#> * Monotonic relationship: warning: Relationship may be non-monotonic. (method=Spearman
↪  correlation; statistic=793638.65; p=0.014)
#> * 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.
#>
```

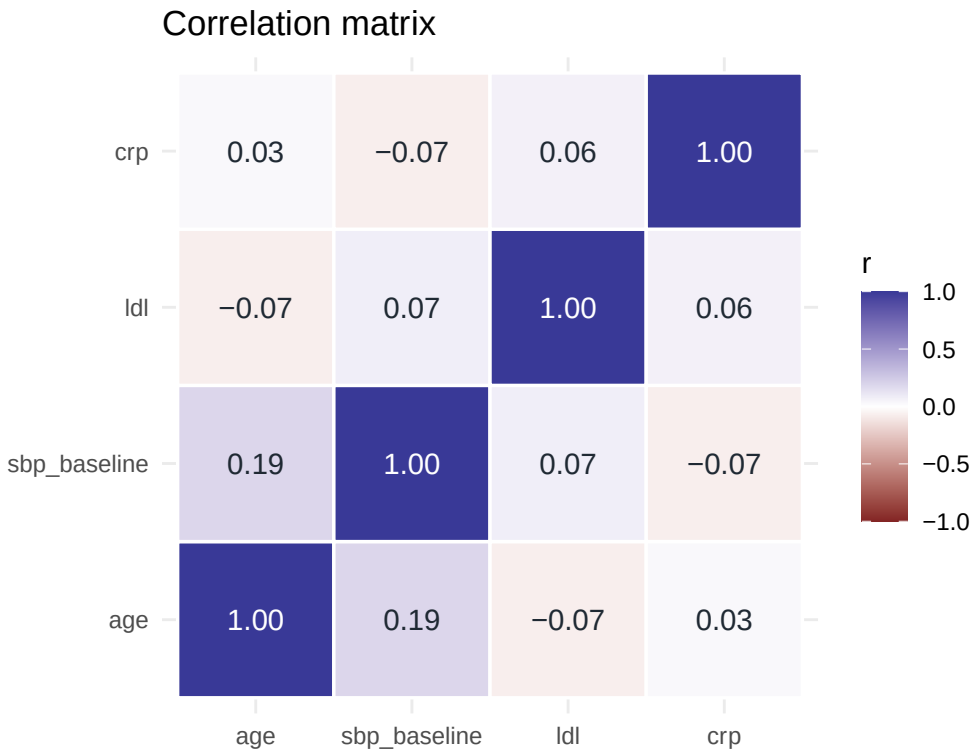
```
#> Recommended test
#> Spearman Correlation
#>
#> Result
#> H0: the correlation between age and sbp_3m is 0.
#> statistic = 793638.65, p = 0.014
#>
#> Effect size
#> Spearman Correlation r: 0.18, small
#>
#> Method comparison
#>    method statistic p.value
#>   Pearson      0.15   0.041
#>  Spearman      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 Correlation, statistic = 793638.65, p = 0.014. The effect size
↪  was small (Spearman Correlation r = 0.18). H0: the correlation between age and sbp_3m
↪  is 0.
plot(x_correlation)
#> `geom_smooth()` using formula = 'y ~ x'
```



`test_correlation_matrix()` extends this to all pairwise correlations among 3+ numeric variables, using pairwise-complete observations per pair and Benjamini-Hochberg-adjusted p-values across all pairs.

```
x_corr_matrix <- test_correlation_matrix(cardio, c(age, sbp_baseline, ldl, crp), method =
↪  "spearman")
#> Warning: `correlation_matrix` is a screening workflow, not a single hypothesis
```

```
#> test.
x_corr_matrix
#> Correlation Matrix
#>
#> Outcome: age, sbp_baseline, ldl, crp
#>
#> Assumptions
#> * Variable types: acceptable: All selected variables should be numeric or ordinal.
#> * Pairwise missing data: acceptable: Pairwise complete observations are used for each
↪  correlation.
#> * Multiple testing correction: acceptable: 0 of 6 pairwise correlation(s) remain
↪  significant after Benjamini-Hochberg adjustment (see the `p.adj` column).
↪  (method=Benjamini-Hochberg (BH))
#>
#> Recommended test
#> Spearman Correlation Matrix
#>
#> Result
#> pairwise correlations reported; smallest pairwise p = 0.012
#>
#> Effect size
#> Maximum absolute r: 0.19, small
#>
#> Pairwise correlations
#>          var1         var2 estimate     p p.adj
#>           age sbp_baseline     0.19 0.012 0.072
#>           age          ldl    -0.07 0.362 0.470
#>           age          crp     0.03 0.712 0.712
#>  sbp_baseline          ldl     0.07 0.333 0.470
#>  sbp_baseline          crp    -0.07 0.343 0.470
#>           ldl          crp     0.06 0.392 0.470
#>
#> Report
#> The correlation matrix workflow for age, sbp_baseline, ldl, crp reported pairwise
↪  correlations using Spearman Correlation Matrix; the smallest pairwise p-value was
↪  0.012. The strongest absolute correlation was 0.19 (small).
plot(x_corr_matrix)
```



Regression

Linear regression

`test_linear_regression()` fits ordinary least squares:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

and reports the overall model F-test as the primary result:

$$F = \frac{(SS_{total} - SS_{res})/p}{SS_{res}/(n-p-1)} \sim F_{p,\,n-p-1}$$

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

Per-term coefficients with confidence intervals are in `alternative_tests$coefficients`. Residual normality (Shapiro-Wilk), homoscedasticity (Breusch-Pagan score test), and multicollinearity (VIF, when $p > 1$) are checked as assumptions.

```
x_linreg <- test_linear_regression(sbp_3m ~ age + ldl, data = cardio)
x_linreg
#> 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; statistic=0.99; p=0.264)
#> * Homoscedasticity: acceptable: Residual variance looks constant across fitted values.
↪  (method=Breusch-Pagan (score test); statistic=0.37; p=0.545)
#> * Multicollinearity: acceptable: Variance inflation factors are all below 5.
↪  (method=Variance inflation factor; statistic=1.01; Max VIF (age) = 1.01)
#>
#> 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 using Linear regression, statistic = 2.91, df = 2.00, p = 0.057.
↪  The effect size was small (R squared = 0.03). H0: none of age, ldl explain variation
↪  in sbp_3m (all slopes equal 0).
plot(x_linreg)
```



Logistic regression

`test_logistic_regression()` fits:

$$\log \frac{P(y_i = 1)}{1 - P(y_i = 1)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

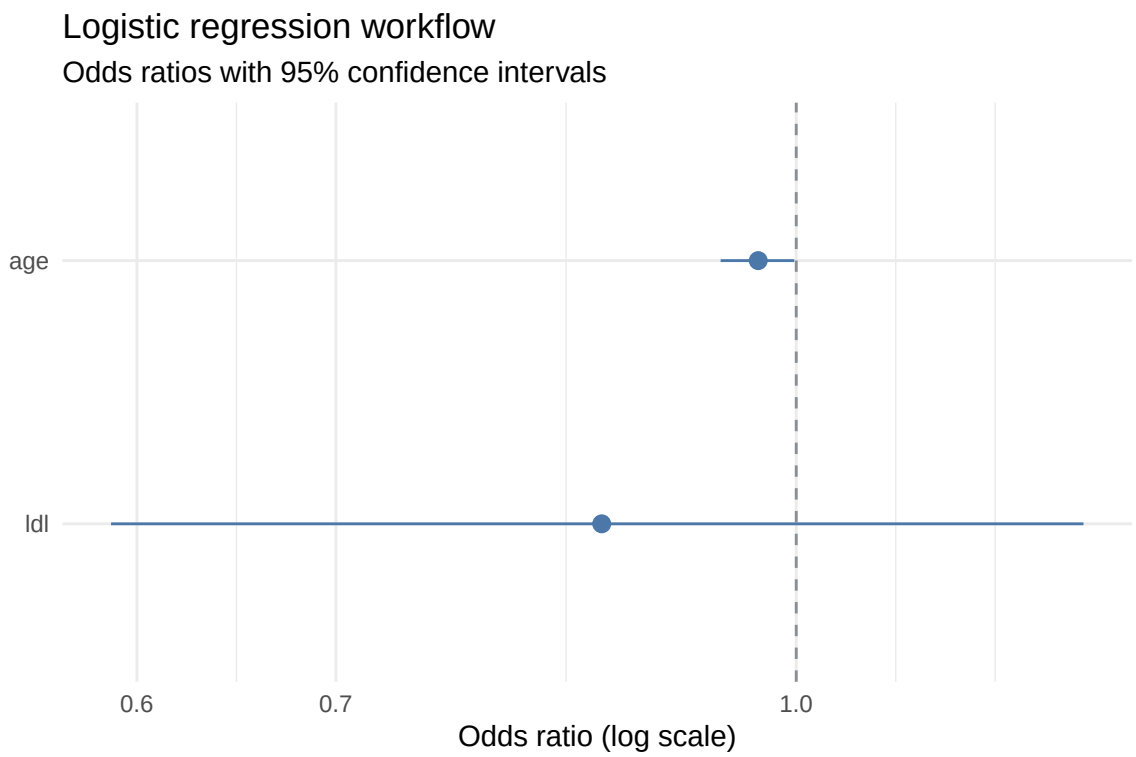
with the overall likelihood-ratio test against the intercept-only model as the primary result:

$$LR = D_{null} - D_{model} \sim \chi^2_p, \qquad D = -2 \log L$$

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

Coefficients are reported on both the log-odds and odds-ratio ($OR_j = e^{\beta_j}$) scales. Multicollinearity (VIF) and influential observations (Cook's distance, flagged at $> 4/n$) are checked as assumptions.

```
x_logreg <- test_logistic_regression(controlled_3m ~ age + ldl, data = cardio)
x_logreg
#> 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 factor; statistic=1.01; Max VIF (age) = 1.01)
#> * Influential observations: warning: 5 observation(s) exceed the Cook's distance
↪  threshold (4/n); consider inspecting them. (method=Cook's distance; statistic=0.05;
↪  Threshold = 0.0222)
#>
#> 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 using Logistic regression, statistic = 4.65, df = 2.00, p = 0.098.
↪  The effect size was small (McFadden's pseudo R squared = 0.02). H0: none of age, ldl
↪  are associated with controlled_3m (all log-odds slopes equal 0).
plot(x_logreg)
```



Survival analysis

`make_cardio_data()` has no time-to-event columns, so this section uses a small synthetic example. `Surv()` is re-exported from the `survival` package.

```
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)
)
```

Log-rank test

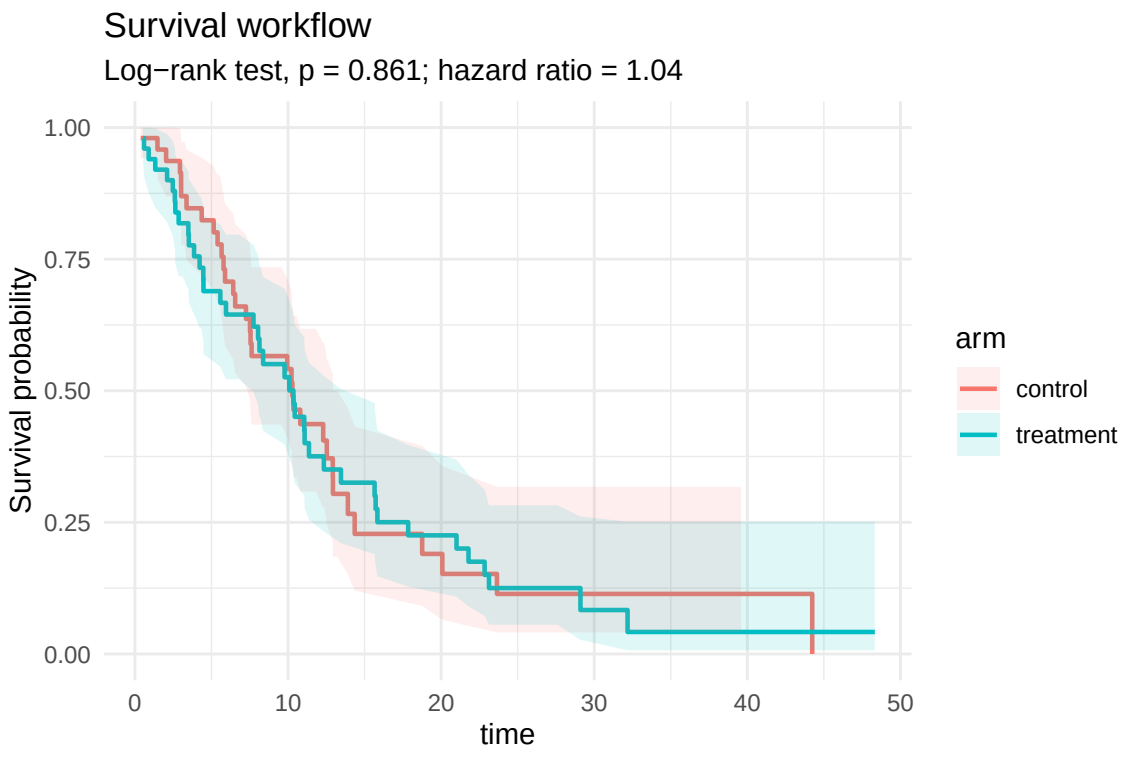
`test_survival()` compares two Kaplan-Meier survival curves via the log-rank test:

$$\chi^2 = \sum_{g \in \{A,B\}} \frac{(O_g - E_g)^2}{V_g} \sim \chi^2_1 \qquad \text{(Mantel, 1966)}$$

with a companion univariate Cox hazard ratio $HR = e^{\hat{\beta}}$ reported 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.

```
x_survival <- test_survival(Surv(time, status) ~ arm, data = survival_dat)
x_survival
#> 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 Log-rank test, statistic = 0.03, df = 1.00, p = 0.861. The
↪  95% confidence interval was [0.66, 1.65]. The effect size (Hazard ratio) was 1.04.
↪  H0: the survival distributions are equal across levels of arm.
plot(x_survival)
#> Warning: Removed 1 row containing missing values or values outside the scale range
#> (`geom_ribbon()`).
```



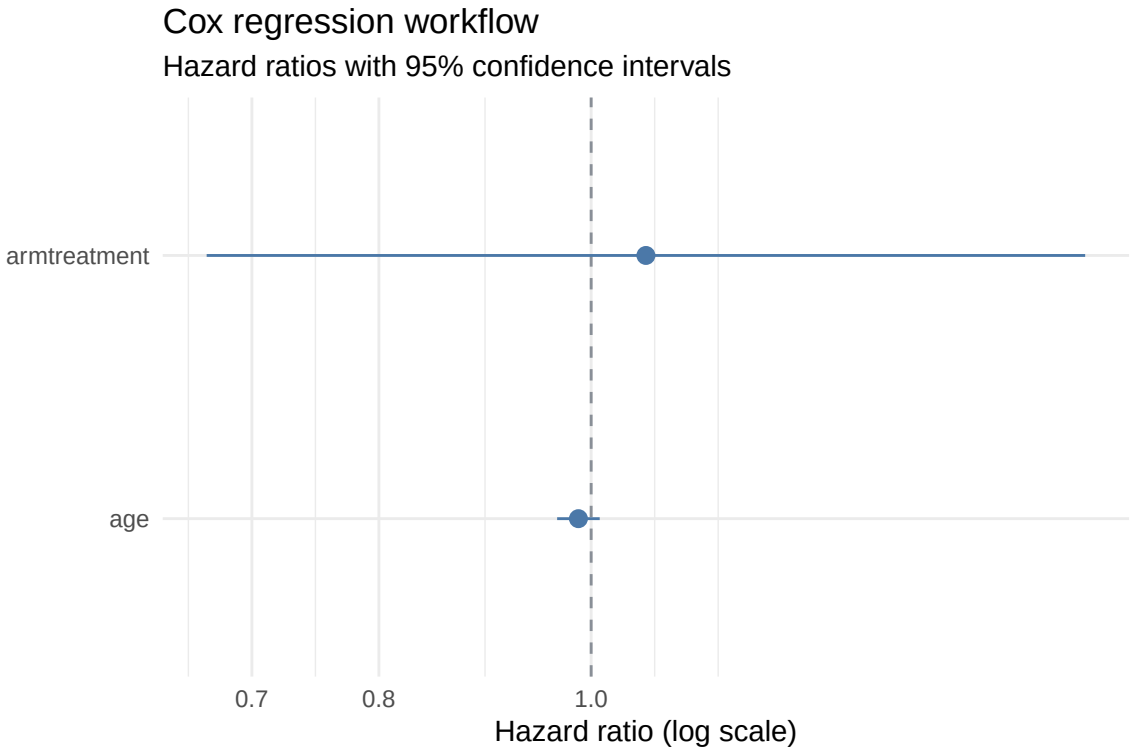
Cox proportional hazards regression

`test_cox()` leaves the baseline hazard $h_0(t)$ unspecified:

$$h(t \mid x) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$

Coefficients are reported as hazard ratios $HR_j = e^{\beta_j}$; the primary result is the overall likelihood-ratio test on p degrees of freedom. Effect size is Harrell's concordance index. The proportional-hazards assumption is checked via the correlation between scaled Schoenfeld residuals and time (Grambsch & Therneau, 1994).

```
x_cox <- test_cox(Surv(time, status) ~ age + arm, data = survival_dat)
x_cox
#> 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 residual test (Grambsch-Therneau); statistic=0.61; p=0.735)
#>
#> 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 result using Cox proportional hazards regression, statistic = 1.42, df =
↪  2.00, p = 0.491. The effect size (Concordance index) was 0.55. H0: none of age, arm
↪  are associated with the hazard (all log hazard ratios equal 0).
plot(x_cox)
```



Diagnostic test accuracy and agreement

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

Diagnostic accuracy

With true/false positives/negatives TP, FP, FN, TN from a 2x2 test-vs. -reference table:

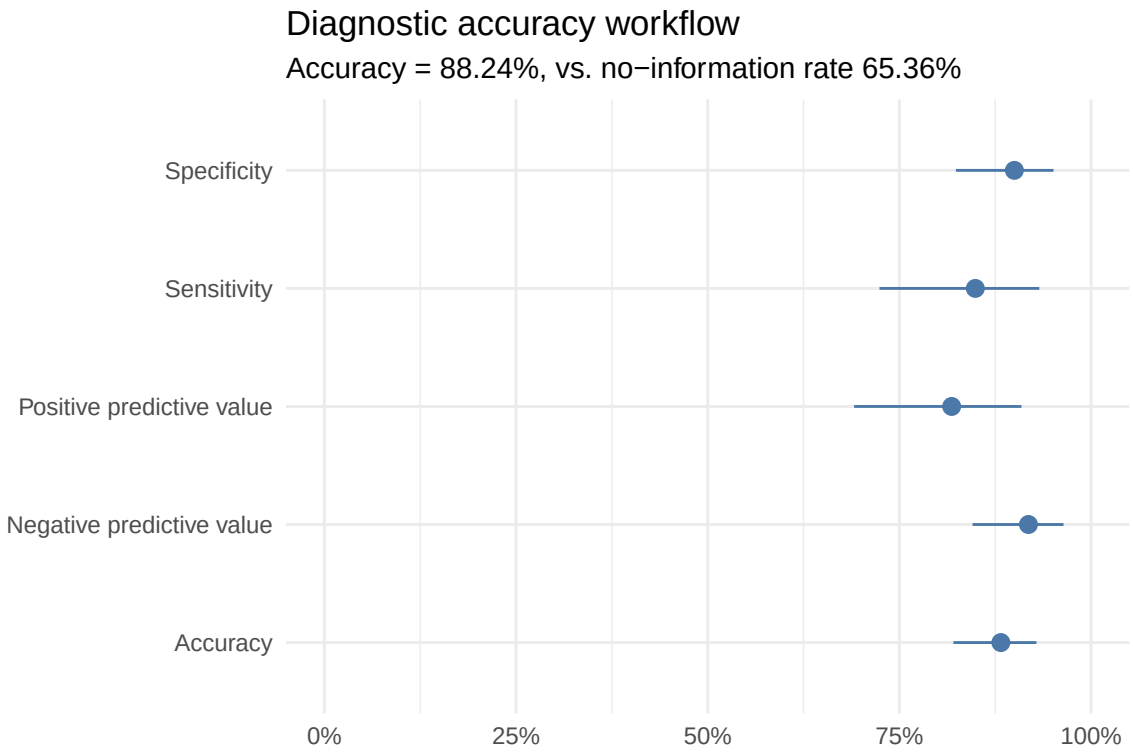
$$Sensitivity = \frac{TP}{TP+FN}, \quad Specificity = \frac{TN}{TN+FP}, \quad PPV = \frac{TP}{TP+FP}, \quad NPV = \frac{TN}{TN+FN}$$

$$LR^{+} = \frac{Sensitivity}{1-Specificity}, \qquad LR^{-} = \frac{1-Sensitivity}{Specificity}$$

Sensitivity, specificity, PPV, NPV, and accuracy each get an exact (Clopper-Pearson) confidence interval; likelihood ratios use the closed-form log-scale interval of Simel, Samsa & Matchar (1991), $SE[\log LR^{+}] = \sqrt{1/TP - 1/(TP+FN) + 1/FP - 1/(FP+TN)}$. `test_diagnostic()` tests overall accuracy against the no-information rate (the larger reference-class proportion) as the primary result, matching `caret::confusionMatrix()`'s convention.

```
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))
)
x_diagnostic <- test_diagnostic(diag_dat, test, reference)
```

```
x_diagnostic
#> Diagnostic Test Accuracy
#>
#> Test: test
#> Reference: reference
#>
#> Assumptions
#> * Independence of observations: assumed: Test and reference results are assumed to be
↪  measured independently.
#>
#> 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 Diagnostic accuracy (2x2 table), statistic = 135.00, df = 153.00, p =
↪  <0.001. The 95% confidence interval was [0.83, 1.00]. The effect size (Accuracy) was
↪  0.88. H0: accuracy equals the no-information rate (0.65).
plot(x_diagnostic)
```



ROC / AUC

`test_roc()` computes the AUC from the Mann-Whitney U relationship (the probability a random positive-class observation outranks a random negative-class one):

$$AUC = \frac{U}{n_1 n_0}$$

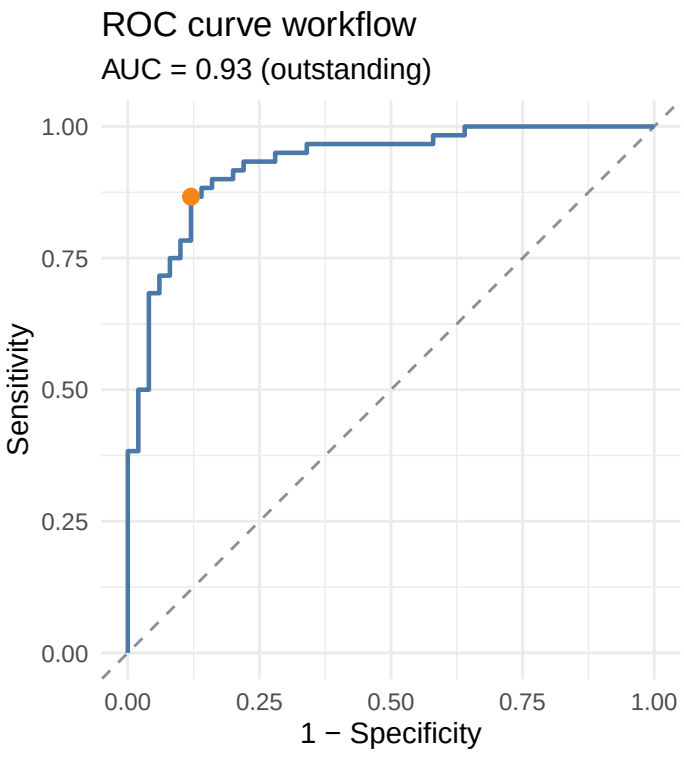
with the closed-form Hanley & McNeil (1982) standard error:

$$SE(AUC) = \sqrt{\frac{AUC(1-AUC) + (n_1-1)(Q_1 - AUC^2) + (n_0-1)(Q_2 - AUC^2)}{n_1 n_0}}, \quad Q_1 = \frac{AUC}{2 - AUC}, \; Q_2 = \frac{2AUC^2}{1 + AUC}$$

The optimal threshold maximizes Youden's J (1950), $J = Sensitivity + Specificity - 1$.

```
roc_dat <- tibble::tibble(
  marker = c(rnorm(60, 2, 1), rnorm(50, 0, 1)),
  disease = c(rep("yes", 60), rep("no", 50))
)
x_roc <- test_roc(roc_dat, marker, disease)
x_roc
#> ROC Curve
#>
#> Predictor: marker
#> Outcome: disease
#>
#> Assumptions
#> * Independence of observations: assumed: Higher predictor values are assumed to
↪  indicate the positive class.
```

```
#>
#> 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, statistic = 17.24, p = <0.001. The 95% confidence interval was [0.88,
↪  0.98]. The effect size was outstanding (AUC = 0.93). H0: AUC = 0.5 (marker does not
↪  discriminate between levels of disease).
plot(x_roc)
```



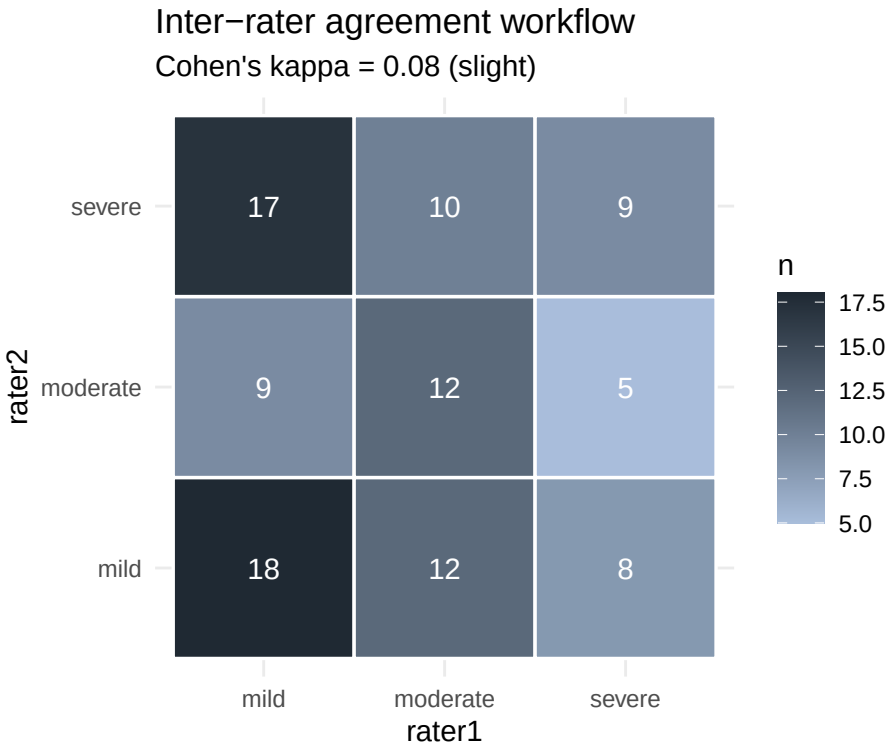
Inter-rater agreement (Cohen's kappa)

`test_agreement()` reports Cohen's kappa from observed agreement p_o and chance-expected agreement p_e:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Two *different* standard errors are used for two different purposes: the confidence interval uses the large-sample SE of Fleiss, Cohen & Everitt (1969), evaluated at $\hat{\kappa}$; the z-test of $H_0 : \kappa = 0$ uses the smaller-magnitude null-hypothesis SE of Fleiss (1981), evaluated at $\kappa = 0$. Reusing the CI SE for the null test (an error `testflow` previously made and has since fixed) understates the null SE whenever agreement is materially above chance, inflating the z-statistic and anti-conservatively shrinking the p-value.

```
agree_dat <- tibble::tibble(
  rater1 = sample(c("mild", "moderate", "severe"), 100, replace = TRUE),
  rater2 = sample(c("mild", "moderate", "severe"), 100, replace = TRUE)
)
x_agreement <- test_agreement(agree_dat, rater1, rater2)
x_agreement
#> 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 = 1.20, p = 0.229, 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 Cohen's kappa, statistic = 1.20, p = 0.229. The 95% confidence interval
↪  was [-0.06, 0.22]. The effect size was slight (Cohen's kappa = 0.08). H0: agreement
↪  beyond chance is zero (kappa = 0).
plot(x_agreement)
```



Intraclass correlation coefficient

`test_icc()` reports three ICC forms from one-way and two-way ANOVA variance decompositions (Shrout & Fleiss, 1979) of n subjects by k raters:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

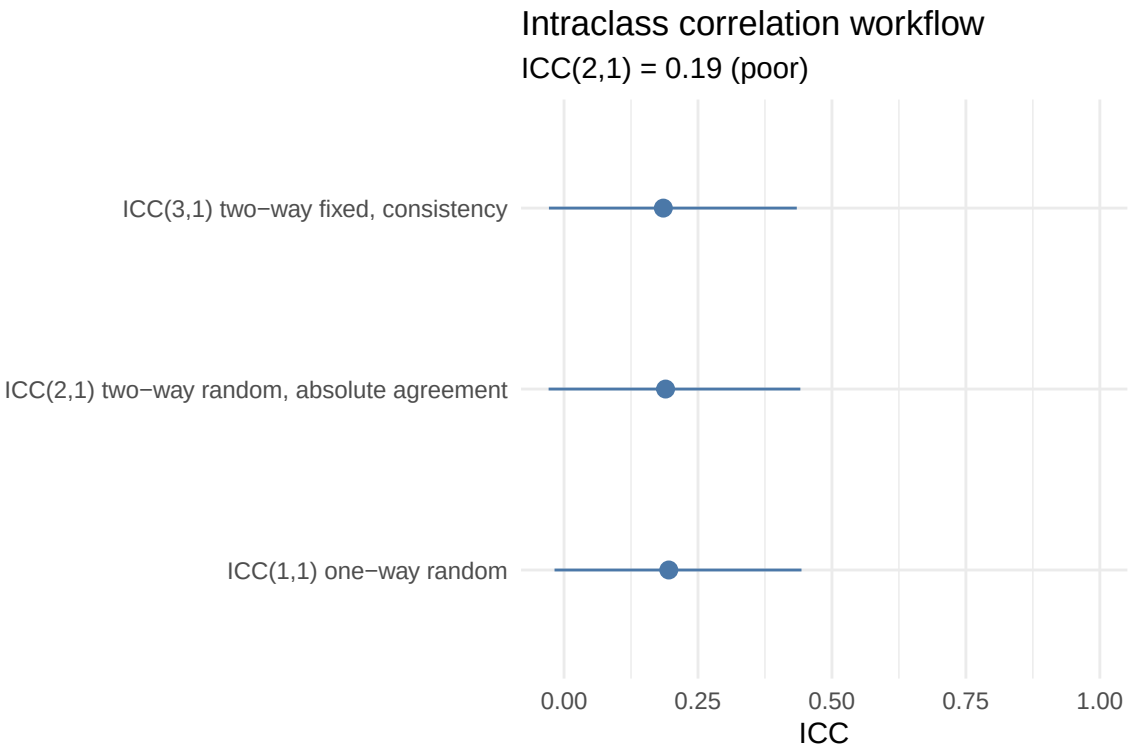
$$ICC(2,1) = \frac{BMS - EMS}{BMS + (k-1)EMS + \frac{k}{n}(JMS - EMS)} \qquad \text{(two-way random, absolute agreement)}$$

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1)EMS} \qquad \text{(two-way fixed, consistency)}$$

ICC(2,1) is reported as the primary/effect-size estimate, following Koo & Li's (2016) recommendation for typical reliability studies. Confidence intervals use the F-distribution formulas of McGraw & Wong (1996), with a Satterthwaite-approximated denominator degrees of freedom for ICC(2,1).

```
icc_dat <- tibble::tibble(
  rater1 = rnorm(30, 50, 10),
  rater2 = rnorm(30, 50, 10),
  rater3 = rnorm(30, 50, 10)
)
x_icc <- test_icc(icc_dat, c(rater1, rater2, rater3))
x_icc
#> 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 using Intraclass correlation coefficient, statistic = 1.68, df =
↪  29.00, p = 0.046. The 95% confidence interval was [-0.03, 0.44]. The effect size was
↪  poor (ICC(2,1) = 0.19). H0: ICC(2,1) = 0 (no reliability beyond chance).
plot(x_icc)
```



Outlier screening

`test_outliers()` is a screening workflow, not a hypothesis test: it flags rows via the IQR rule (univariate, per variable),

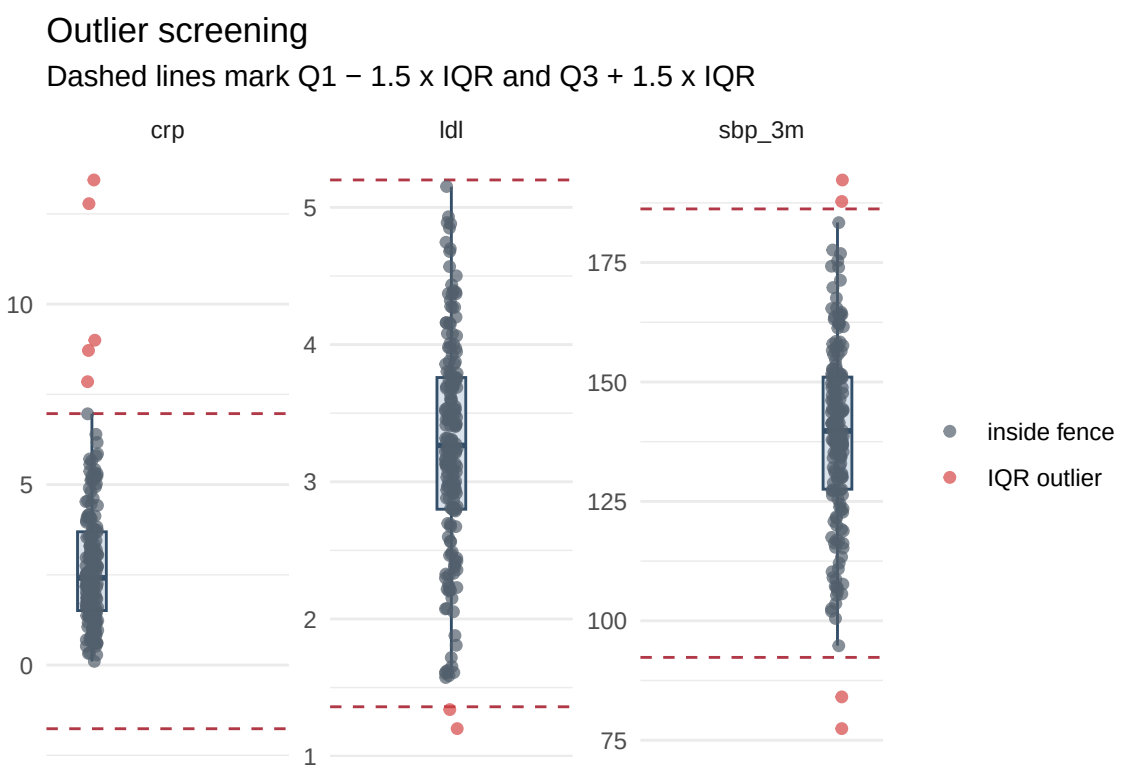
$$x < Q_1 - 1.5\,IQR \quad \text{or} \quad x > Q_3 + 1.5\,IQR$$

and/or Mahalanobis distance (multivariate, `method = "mahalanobis"`), flagged against a χ^2 cutoff at the 97.5th percentile with degrees of freedom equal to the number of variables:

$$D_i^2 = (x_i - \bar{x})^\top S^{-1} (x_i - \bar{x}), \qquad \text{flag if } D_i^2 > \chi^2_{0.975,\,p}$$

```
x_outliers <- test_outliers(c(sbp_3m, ldl, crp), data = cardio)
#> Warning: `outliers` is a screening workflow, not a single hypothesis test.
x_outliers
#> 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 skewed.
#>
#> Recommended test
#> IQR outlier detection
#>
```

```
#> Result
#> flagged rows = 11
#>
#> Effect size
#> Outlier count: 11.00
#>
#> Report
#> The outlier workflow flagged 11 rows for review.
plot(x_outliers)
```



Effect-size reference

The table below consolidates every effect-size formula reported across the workflows above, since some functions (Kendall's W, discordant pairs, observed proportion) are only summarized inline elsewhere.

Workflow	Effect size	Formula
One sample	Cohen's d	$d = (\bar{x} - \mu_0)/s_x$
Two groups (t-test)	Cohen's d	$d = (\bar{x}_1 - \bar{x}_2)/s_p$
Two groups (Wilcoxon)	Rank-biserial r	$r_{rb} = 1 - 2W/(n_1 n_2)$
Paired	Cohen's dz	$d_z = \bar{d}/s_d$
ANOVA (one-way, factorial)	Eta squared	$\eta_j^2 = SS_j / \sum SS$
Kruskal-Wallis	Epsilon squared	$\epsilon^2 = (H - k + 1)/(n - k)$
Repeated ANOVA	Eta squared (time)	$\eta_{time}^2 = SS_{time}/SS_{total}$
Friedman	Kendall's W	$W = \chi_F^2/(n(k-1))$
Cochran Q	Kendall's-W analogue	$W = Q/(n(k-1))$
Categorical association	Cramer's V	$V = \sqrt{\chi^2/(n(\min(r,c)-1))}$
One-sample proportion	Observed proportion	$\hat{p} = x/n$

Workflow	Effect size	Formula
Multinomial	Chi-square GOF statistic	$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$
Paired categorical	Discordant pairs	$b + c$
Correlation	r / rho / tau	see "Correlation" above
Linear regression	R squared, adjusted R squared	$R^2 = 1 - SS_{res}/SS_{total}$
Logistic regression	McFadden's pseudo R squared	$1 - \log L_{model} / \log L_{null}$
Survival (log-rank)	Hazard ratio	$HR = e^{\hat{\beta}}$ (no magnitude label)
Cox regression	Concordance index	Harrell's C (no magnitude label)
Diagnostic accuracy	Accuracy	$(TP + TN)/n$ (no magnitude label)
ROC	AUC	$U/(n_1 n_0)$
Agreement	Cohen's kappa	$(p_o - p_e)/(1 - p_e)$
ICC	ICC(2,1)	see "Intraclass correlation" above

Magnitude labels

- Cohen's d: negligible < 0.2, small < 0.5, moderate < 0.8, otherwise large.
- Eta/epsilon/R-squared family: negligible < 0.01, small < 0.06, moderate < 0.14, otherwise large. (McFadden's pseudo R^2 is not on the OLS R^2 scale; values above ~0.2-0.4 are already a good fit in practice.)
- Cramer's V, rank-biserial $|r|$, Kendall's W: negligible < 0.1, small < 0.3, moderate < 0.5, otherwise large.
- AUC: poor if $\max(AUC, 1 - AUC) < 0.7$, acceptable < 0.8, excellent < 0.9, otherwise outstanding (Hosmer, Lemeshow & Sturdivant 2013, p. 177).
- Cohen's kappa: poor < 0, slight < 0.21, fair < 0.41, moderate < 0.61, substantial < 0.81, otherwise almost perfect (Landis & Koch, 1977).
- ICC: poor < 0.5, moderate < 0.75, good < 0.9, otherwise excellent (Koo & Li, 2016).
- Hazard ratio, concordance index, accuracy: no magnitude label - there is no widely agreed small/moderate/large convention, and inventing one would overstate how standardized that judgment is. The estimate itself is always reported.

Sample-size planning

`sample_size()` and the endpoint-specific `sample_size_*()` functions plan continuous, binary, survival, ordinal, bioequivalence, and precision-based designs, each under superiority, non-inferiority, or equivalence objectives where applicable. Every result is a `sample_size` object with a raw and dropout-adjusted n, the assumptions used, a report-ready sentence, and (for some designs) a power curve.

Notation: α is the type I error rate, $1 - \beta$ is power, $z_{1-\alpha}/z_{1-\alpha/2}/z_{1-\beta}$ are standard normal quantiles, $r = n_B/n_A$ is the allocation ratio. Superiority uses two-sided α; non-inferiority/equivalence use one-sided α (pass `alpha = 0.025` explicitly - `testflow` does not rescale it). Dropout adjustment is always:

$$n_{recruit} = \left\lceil \frac{n_{eval}}{1 - d} \right\rceil$$

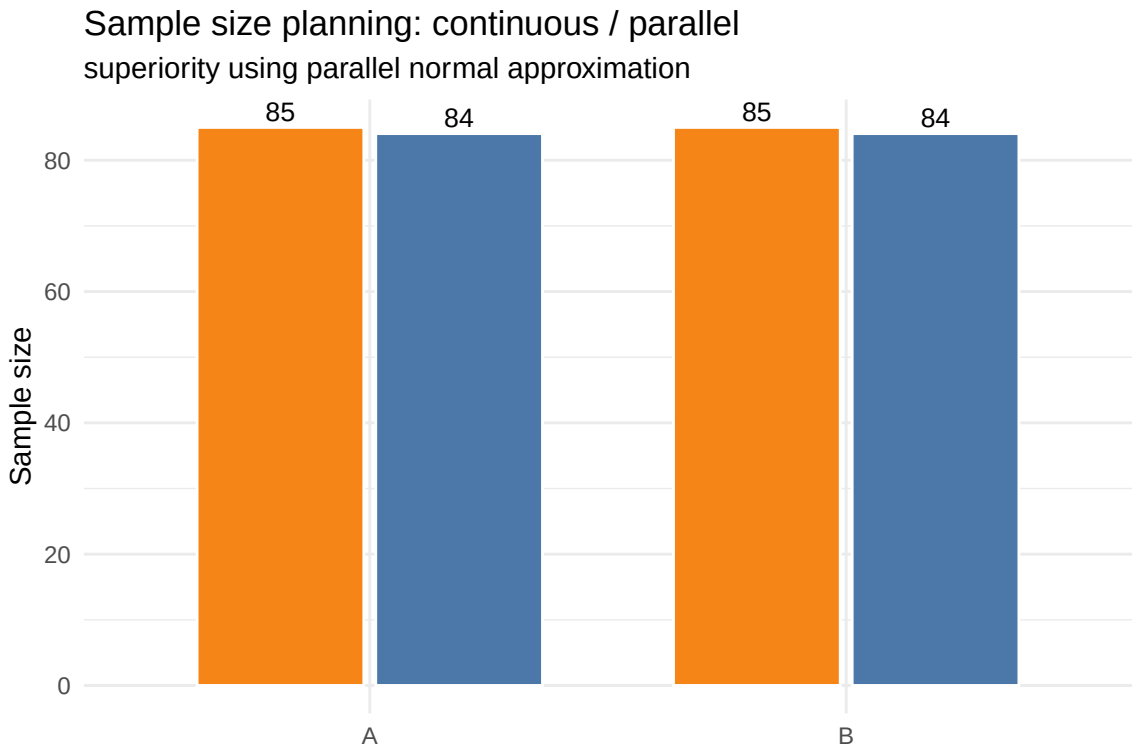
```
sample_size(
  endpoint = c("continuous", "binary", "survival", "ordinal"),
  design = c("parallel", "paired", "repeated"),
  objective = c("superiority", "noninferiority", "equivalence"),
  ...
)
```

Continuous endpoints

`sample_size_continuous()` covers parallel, paired, and repeated designs. For a parallel-group superiority comparison with common variance σ^2 and target difference Δ:

$$n_A = \frac{(r+1)\sigma^2(z_{1-\beta} + z_{1-\alpha/2})^2}{r\Delta^2}, \qquad n_B = rn_A$$

```
ss_parallel <- sample_size_continuous(design = "parallel", objective = "superiority",
↪  delta = 5, sd = 10, alpha = 0.05, power = 0.90)
ss_parallel
#> Sample size planning
#>
#> Endpoint: continuous
#> Design: parallel
#> Objective: superiority
#> Method: parallel normal approximation
#>
#> Assumptions
#> * Planning note 1: assumed: Two independent groups are assumed.
#> * Planning note 2: assumed: Allocation ratio B/A = 1.000.
#>
#> Result
#> Raw n = 84.059
#> Adjusted n = A=85, B=85
#> Adjusted per-group n = A=85, B=85
#>
#> Report
#> Parallel continuous superiority sample size was calculated with delta = 5.000, sd =
↪  10.000, allocation ratio = 1.000; required per-group sizes = A 85, B 85.
plot(ss_parallel, type = "summary")
```



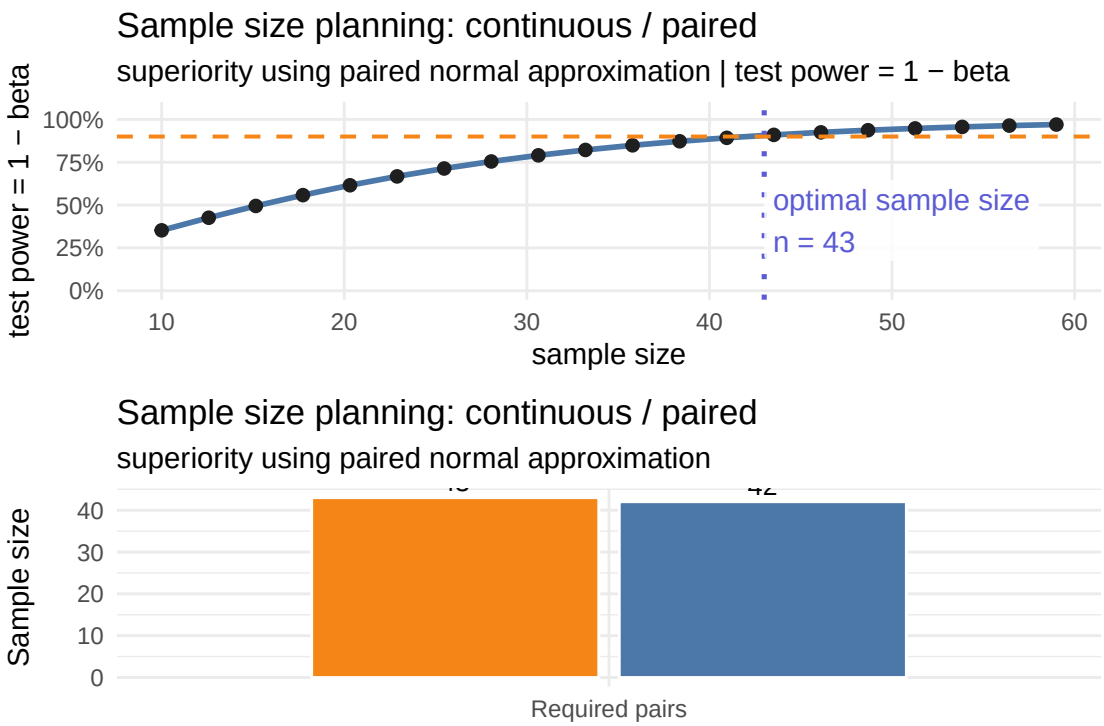
Paired designs replace σ with the paired-difference SD and drop the allocation ratio:

$$n_{pairs} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_{diff}^2}{\Delta^2}$$

`curve_data` (and therefore `plot(type = "curve"/"both")`) is populated only for `design = "paired"`/`"repeated"`, `objective = "superiority"`; other combinations fall back to the summary bar chart shown above.

```
ss_paired <- sample_size_continuous(design = "paired", objective = "superiority", delta =
↪  5, sd_diff = 10, alpha = 0.05, power = 0.90)
ss_paired
#> Sample size planning
#>
#> Endpoint: continuous
#> Design: paired
#> Objective: superiority
#> Method: paired normal approximation
#>
#> Assumptions
#> * Planning note 1: assumed: Paired observations are assumed.
#> * Planning note 2: assumed: Effective SD = 10.000.
#>
#> Result
#> Raw n = 42.030
#> Adjusted n = 43
#> Adjusted total n = 43
#>
#> Report
#> Paired continuous superiority sample size was calculated with delta = 5.000, sd_diff =
↪  10.000, alpha = 0.050, power = 0.900; required pairs = 43.
```

```r
plot(ss_paired, type = "both")
```



For `design = "repeated"` with 3+ time points, σ_{diff} is rescaled under a compound-symmetry working correlation before the paired formula is applied (`testflow` design-effect extension, not a standalone textbook formula):

$$\sigma_{eff} = \sigma_{diff}\sqrt{\frac{1 + (m-1)\rho}{m}}$$

Non-inferiority uses $d_{NI} = \Delta + \delta$; equivalence uses $d_{EQ} = \delta - |\Delta|$ (or, when $\Delta = 0$ exactly, $z_{1-\beta/2}$ in place of $z_{1-\beta}$), both plugged into the same form as superiority with a one-sided α.

Binary endpoints

`sample_size_binary()` covers parallel risk-difference and paired (discordant-pairs) designs. Parallel superiority (`method = "pooled"`, with $\bar{p} = (p_A + r p_B)/(1 + r)$):

$$n_A = \frac{\left[z_{1-\alpha/2}\sqrt{(1 + 1/r)\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_A(1-p_A) + p_B(1-p_B)/r}\right]^2}{(p_A - p_B)^2}$$

```r
sample_size_binary(design = "parallel", objective = "superiority", p1 = 0.4, p2 = 0.25,
↪  method = "pooled", alpha = 0.05, power = 0.90)
#> Sample size planning
#>
#> Endpoint: binary
#> Design: parallel
#> Objective: superiority
#> Method: parallel pooled normal approximation
#>
```

```
#> Assumptions
#> * Planning note 1: assumed: Two independent groups are assumed.
#> * Planning note 2: assumed: Risk difference = 0.150.
#>
#> Result
#> Raw n = 202.810
#> Adjusted n = A=203, B=203
#> Adjusted per-group n = A=203, B=203
#>
#> Report
#> Parallel binary superiority sample size was calculated with p1 = 0.400, p2 = 0.250,
↪  allocation ratio = 1.000; required per-group sizes = A 203, B 203.
```

Paired binary designs plan on the discordant cells directly, with discordant odds ratio $OR_D = p_{10}/p_{01}$ and discordance rate $\lambda_D = p_{10} + p_{01}$:

$$n_{total} = \left\lceil \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (OR_D + 1)^2}{(OR_D - 1)^2 \lambda_D} \right\rceil$$

Non-inferiority/equivalence use the same asymmetric variance form at the non-inferiority/equivalence distance instead of $p_A - p_B$.

Survival endpoints

`sample_size_survival()` plans required events, then converts to total N. Under an assumed exponential hazard:

$$D_{total} = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{[\log(HR)]^2}$$

or, proportional-hazards-only (`method = "ph_only"`):

$$D_{total} = \frac{2(HR + 1)^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{(HR - 1)^2}$$

Given planned survival probabilities $S_A(t)$, $S_B(t)$, equal allocation:

$$N_{total} = \frac{2D_{total}}{(1 - S_A(t)) + (1 - S_B(t))}$$

```
sample_size_survival(hr = 0.7, survival_a = 0.8, survival_b = 0.7, alpha = 0.05, power =
↪  0.90)
#> Sample size planning
#>
#> Endpoint: survival
#> Design: parallel
#> Objective: superiority
#> Method: survival exponential event approximation
#>
#> Assumptions
#> * Planning note 1: assumed: Two independent groups are assumed.
#> * Planning note 2: assumed: Event probabilities are derived from the planned survival
↪  probabilities.
```

```
#>
#> Result
#> Raw n = 1321.512
#> Adjusted n = 1,322
#> Adjusted total n = 1,322
#> Required events = 331
#>
#> Report
#> Survival superiority sample size was calculated with hr = 0.700; required total events
↪  = 331, estimated total sample size = 1,322.
```

With staggered uniform accrual (`accrual_duration`, `follow_up`), the flat $1 - S_g(t)$ is replaced by the accrual-averaged event probability:

$$\bar{q}_g = 1 - \frac{e^{-\lambda_g (T-R)} - e^{-\lambda_g T}}{\lambda_g R}, \qquad T = R + \text{follow_up}$$

which always requires more subjects than instantaneous accrual for the same event target, and reduces to the flat conversion as $R \to 0$. Non-inferiority/equivalence work on the log-hazard-ratio scale with $d_{NI} = \log(HR_M) - \log(HR)$ or $d_{EQ} = \min(\theta - \theta_L, \theta_U - \theta)$.

Ordinal endpoints

`sample_size_ordinal()` implements Noether's method for a two-group Mann-Whitney/Wilcoxon comparison, with $P = P(A > B) + \frac{1}{2}P(A = B)$:

$$n_{per\ group} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{6(P - 0.5)^2}$$

```
sample_size_ordinal(p_superiority = 0.65, alpha = 0.05, power = 0.90)
#> Sample size planning
#>
#> Endpoint: ordinal
#> Design: parallel
#> Objective: superiority
#> Method: Noether superiority approximation
#>
#> Assumptions
#> * Planning note 1: assumed: Two independent ordinal groups are assumed.
#>
#> Result
#> Raw n = 77.833
#> Adjusted n = A=78, B=78
#> Adjusted per-group n = A=78, B=78
#> Adjusted total n = 156
#>
#> Report
#> Ordinal superiority sample size was calculated with p_superiority = 0.650; required
↪  per-group size = 78, total sample size = 156.
```

Bioequivalence

`sample_size_bioequivalence()` plans a two one-sided test (TOST) study on the log scale, with $\theta_0 = \log(GMR)$, bounds $\theta_L = \log(L)/\theta_U = \log(U)$ (usually 0.80/1.25), and $d_{BE} = \min(\theta_0 - \theta_L, \theta_U - \theta_0)$. The default `method = "iterative_tost"` searches for the smallest n achieving the *exact* TOST power via the noncentral t distribution (Phillips 1990):

$$Power(n) = T_{df,\delta_U}(-t_{1-\alpha,df}) - T_{df,\delta_L}(t_{1-\alpha,df})$$

where $T_{df,\delta}$ is the noncentral t CDF, $\delta_U = (\theta_0 - \theta_U)/SE$, $\delta_L = (\theta_0 - \theta_L)/SE$, $\sigma = \sqrt{\log(1 + CV^2)}$. Verified in package tests to match `PowerTOST::power.TOST()` (the regulatory-standard, Owen's-Q-based calculation) for balanced designs; `method = "normal_approx"` gives the closed-form approximation $n_{total} \approx 2\sigma_w^2(z_{1-\beta} + z_{1-\alpha})^2/d_{BE}^2$, which agrees closely near $GMR = 1$ but can diverge further away from it (an earlier version of `testflow` used only the normal approximation here, under-sizing studies by up to ~20% at small n).

```
sample_size_bioequivalence(design = "crossover", gmr = 0.95, cv_within = 0.30, alpha =
↪  0.05, power = 0.90)
#> Sample size planning
#>
#> Endpoint: bioequivalence
#> Design: crossover
#> Objective: equivalence
#> Method: crossover bioequivalence (iterative_tost)
#>
#> Assumptions
#> * Planning note 1: assumed: A two-period two-treatment crossover with log-normal
↪  within-subject variation is assumed.
#> * Planning note 2: assumed: Within-subject CV = 0.300.
#> * Planning note 3: assumed: Distance to the nearest bioequivalence boundary (log
↪  scale) = 0.172.
#>
#> Result
#> Raw n = 51.624
#> Adjusted n = 52
#> Adjusted total n = 52
#>
#> Report
#> Crossover bioequivalence sample size was calculated with gmr = 0.950, bounds = [0.800,
↪  1.250]; required total n = 52.
```

Precision-based sample size

`sample_size_precision()` sizes a study to a target confidence-interval half-width w rather than power against an effect - there is no `power` argument.

```
sample_size_precision(endpoint = "continuous", design = "one_sample", width = 2, sd = 10,
↪  alpha = 0.05)
#> Sample size planning
#>
#> Endpoint: continuous
#> Design: one_sample
#> Objective: precision
#> Method: one-sample CI half-width
```

```
#>
#> Assumptions
#> * Planning note 1: assumed: Target CI half-width = 2.000.
#>
#> Result
#> Raw n = 96.036
#> Adjusted n = 97
#> Adjusted total n = 97
#>
#> Report
#> One-sample precision sample size was calculated with sd = 10.000, width = 2.000;
↪  required n = 97.
sample_size_precision(endpoint = "binary", design = "two_sample", width = 0.08, p1 = 0.4,
↪  p2 = 0.3, allocation = 2)
#> Sample size planning
#>
#> Endpoint: binary
#> Design: two_sample
#> Objective: precision
#> Method: two-proportion CI half-width (unequal-allocation risk-difference variance)
#>
#> Assumptions
#> * Planning note 1: assumed: Target CI half-width for the risk difference = 0.080.
#> * Planning note 2: assumed: Allocation ratio n_B / n_A = 2.000; Var(p1_hat - p2_hat) =
↪  p1(1-p1)/n_A + p2(1-p2)/n_B is used directly (this generalizes the equal-allocation
↪  formula, which is the r = 1 special case).
#>
#> Result
#> Raw n = 207.079
#> Adjusted n = A=208, B=415
#> Adjusted per-group n = A=208, B=415
#> Adjusted total n = 623
#>
#> Report
#> Two-proportion precision sample size was calculated with p1 = 0.400, p2 = 0.300, width
↪  = 0.080, allocation ratio = 2.000; required per-group sizes = A 208, B 415.
```

For `endpoint = "binary"`, `design = "one_sample"`, `method` chooses the interval used for planning: `"wald"` (default, closed-form normal approximation) or `"wilson"`/`"exact"` (Clopper-Pearson), which run a deterministic search for the minimum integer n whose interval has maximum one-sided half-width $\leq w$ - important for rare-prevalence studies (newborn screening, rare adverse events) where the Wald approximation is unreliable:

```
sample_size_precision(endpoint = "binary", design = "one_sample", width = 0.02, p = 0.01,
↪  method = "wilson")
#> Warning: Expected event or non-event count is below min_expected_events = 5;
#> rare-event asymptotics may be unreliable regardless of the confidence-interval
#> method.
#> Sample size planning
#>
#> Endpoint: binary
#> Design: one_sample
#> Objective: precision
#> Method: one-proportion CI half-width (wilson, expected criterion)
#>
```

```
#> Assumptions
#> * Planning note 1: assumed: Confidence interval method: Wilson score.
#> * Planning note 2: assumed: if (identical(method, "wilson")) "The interval is
↪  evaluated at the anticipated proportion p itself (no rounding of the event count)."
↪  else "Event counts are evaluated at the single anticipated event count round(n * p)."
#> * Planning note 3: assumed: "Half-width" is the maximum one-sided distance from the
↪  reference proportion to either confidence limit; for asymmetric (Wilson/exact)
↪  intervals this is not the same as half of the total interval width.
#> * Planning note 4: assumed: Wilson/exact planning uses a deterministic integer
↪  numerical search for the minimum sample size (no closed-form formula exists).
#> * Planning note 5: assumed: Precision-based planning targets confidence-interval
↪  width, not the probability of observing any events; a narrow planned interval does
↪  not by itself ensure that cases will be observed (see the rare-event diagnostics).
#>
#> Result
#> Raw n = 280.000
#> Adjusted n = 280
#> Adjusted total n = 280
#>
#> Report
#> One-proportion precision planning used a 95% Wilson score confidence interval,
↪  anticipated prevalence p = 0.010, and requested maximum half-width 0.020. The minimum
↪  complete-case sample size was 280. No dropout inflation was applied; the recruitment
↪  target was 280. The expected number of events is 2.800 and the achieved maximum
↪  half-width is 0.020.
#>
#> Diagnostics
#> Anticipated prevalence: 0.010
#> Confidence level: 0.950
#> CI method: wilson
#> Precision criterion: expected
#> Complete-case n: 280
#> Adjusted (recruitment) n: 280
#> Expected events (adjusted n): 2.800
#> P(zero events): 0.060
#> Achieved maximum half-width: 0.020
#> Note: Expected event or non-event count is below min_expected_events = 5; rare-event
↪  asymptotics may be unreliable regardless of the confidence-interval method.
```

`criterion = "expected"` (default) checks precision at the anticipated event count `round(n*p)`; `criterion = "worst_case"` additionally requires the target to hold across nearby plausible event counts. `x$diagnostics` reports rare-event planning quantities a CI-width target alone does not guarantee (`expected_events`, `probability_zero_events`, `complete_case_n` vs. `adjusted_n`), with an `approximation_warning` when the Wald approximation is used or expected counts are small.

For binary `design = "two_sample"`, with $r = n_B/n_A$:

$$n_A = \frac{z_{1-\alpha/2}^2}{w^2}\left[p_1(1-p_1) + \frac{p_2(1-p_2)}{r}\right], \qquad n_B = r n_A$$

Cluster-randomized adjustment and dropout

`sample_size_cluster_adjust()` inflates an individually randomized n by the design effect $DE = 1+(m-1)\rho$ (equal cluster size m, intracluster correlation ρ), or $DE \approx 1 + ((1 + CV_m^2)m - 1)\rho$ for unevenly sized clusters.

`sample_size_adjust_dropout()` exposes the dropout formula above directly. Both return a plain integer, not a `sample_size` object - they are post-processing adjustments, not full planning calculations.

```
sample_size_cluster_adjust(100, m = 20, rho = 0.02)
#> [1] 138
sample_size_adjust_dropout(100, dropout = 0.15)
#> [1] 118
```

Summary of supported settings

Endpoint	Design(s)	Objective(s)	Formula family
Continuous	parallel, paired, repeated (2+ time points)	superiority, non-inferiority, equivalence	normal-approximation mean-difference formulas, compound-symmetry repeated extension
Binary	parallel, paired, repeated (2 time points)	superiority, non-inferiority, equivalence	risk-difference and discordant-pair planning
Survival	parallel	superiority, non-inferiority, equivalence	event-based proportional-hazards planning, optional uniform-accrual adjustment
Ordinal	parallel	superiority	Noether approximation
Bioequivalence	crossover, parallel	equivalence (TOST)	iterative TOST or normal approximation, log scale
Precision	one-sample, two-sample, log-OR	CI half-width (no power/objective)	continuous and binary CI-width formulas

Every `sample_size` object also has `print()`, `summary()`, `report()`, `as_tibble()`, and `plot(type = "summary"/"curve"/"both")` methods. See `?sample_size_continuous`, `?sample_size_binary`, `?sample_size_survival`, `?sample_size_ordinal`, `?sample_size_bioequivalence`, `?sample_size_precision`, and `?sample_size_cluster_adjust` for the complete argument reference and every `design`/`objective`/`method` combination.

References

Gosset WS (Student). The probable error of a mean. *Biometrika*. 1908;6(1):1-25.

Welch BL. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. 1947;34(1-2):28-35.

Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945;1(6):80-83.

Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*. 1947;18(1):50-60.

Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd; 1925.

Levene H. Robust tests for equality of variances. In: *Contributions to Probability and Statistics*. Stanford University Press; 1960.

Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*. 1952;47(260):583-621.

Tukey JW. Comparing individual means in the analysis of variance. *Biometrics*. 1949;5(2):99-114.

Dunn OJ. Multiple comparisons using rank sums. *Technometrics*. 1964;6(3):241-252.

Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*. 1937;32(200):675-701.

Girden ER. *ANOVA: Repeated Measures*. Sage; 1992.

Cochran WG. The comparison of percentages in matched samples. *Biometrika*. 1950;37(3/4):256-266.

McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153-157.

Pearson K. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*. 1895;58:240-242.

Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*. 1900;50(302):157-175.

Fisher RA. On the interpretation of chi-square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*. 1922;85(1):87-94.

Spearman C. The proof and measurement of association between two things. *The American Journal of Psychology*. 1904;15(1):72-101.

Kendall MG. A new measure of rank correlation. *Biometrika*. 1938;30(1/2):81-93.

Cramer H. *Mathematical Methods of Statistics*. Princeton University Press; 1946.

Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404-413.

Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum; 1988.

Draper NR, Smith H. *Applied Regression Analysis* (3rd ed.). Wiley; 1998.

Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression* (3rd ed.). Wiley; 2013.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958;53(282):457-481.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*. 1966;50(3):163-170.

Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*. 1972;34(2):187-202.

Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.

Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology*. 1991;44(8):763-770.

Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ*. 1994;308(6943):1552.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35.

Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37-46.

Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*. 1969;72(5):323-327.

Fleiss JL. *Statistical Methods for Rates and Proportions* (2nd ed.). Wiley; 1981.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979;86(2):420-428.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1(1):30-46.

Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*. 2016;15(2):155-163.

Julious SA. *Sample Sizes for Clinical Trials*. Chapman & Hall/CRC; 2010.

Phillips KF. Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*. 1990;18(2):137-144.

Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold; 2000.