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)
#> Statistical test workflow
#>
#> Outcome: sbp_3m
#> Design: one numerical sample
#>
#> Assumptions
#> * Independence of observations: assumed: Assumed from study design.
#> * Normality: sbp_3m: acceptable: Approximate normality looks reasonable. (method=Shapiro-Wilk; stati
#> * Symmetry of deviations: not checked: Normality made the symmetry check unnecessary. (method=Signed
#>
#> 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)
#> Statistical test workflow
#>
#> Outcome: sbp_3m
#> Group: sex
#> Design: two independent groups
#>
#> 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)
#> Statistical test workflow
#>
#> Outcome: sbp_3m - sbp_baseline
#> Design: paired measurements
#>
#> 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: not checked: Normality made the symmetry check unnecessary.
#> * 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)
#> Statistical test workflow
#>
#> Outcome: sbp_3m
#> Group: treatment
#> Design: more than two independent groups
#>
```

```
#> 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)
#> Statistical test workflow
#>
#> Outcome: sbp_3m
#> Group: sex, treatment
#> Design: factorial design
#>
#> 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; the workflow still reports the design.
#>
#> Recommended test
#> Factorial ANOVA
#>
#> Result
#> H0: the population mean or location of sbp_3m is equal across levels of sex, treatment.
#> statistic = 3.78, df = 1.00, p = 0.053
#>
#> Effect size
#> eta squared: 0.02, small
#>
#> Report
#> The factorial design workflow for sbp_3m did not show a statistically significant result using Facto
```

Repeated measurements

```
test_repeated(cardio, c(sbp_baseline, sbp_3m, sbp_6m), id = id)
#> Statistical test workflow
#>
#> Outcome: sbp_baseline, sbp_3m, sbp_6m
#> Group: time
#> Design: repeated numeric measurements
#>
#> 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: not checked: Sphericity is not checked here; use this as a teaching note unless a form
#>
#> 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 = 3.76, df = 2.00, p = 0.024
#>
#> Effect size
#> eta squared: 0.05, small
#>
#> 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)
#> Statistical test workflow
#>
#> Outcome: treatment
#> Group: controlled_3m
#> Design: two categorical variables
#>
#> 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))
#> Statistical test workflow
#>
#> Outcome: controlled_baseline, controlled_3m, controlled_6m
#> Design: repeated categorical measurements
#>
#> 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
#>
#> 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*.

Correlation

```
test_correlation(sbp_3m ~ age, data = cardio)
#> Statistical test workflow
#>
#> Outcome: sbp_3m
#> Group: age
#> Design: two numeric variables
#>
#> 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
#>
#> Report
#> The two numeric variables workflow for sbp_3m showed a statistically significant result using Spearm
```

Outliers

```
test_outliers(c(sbp_3m, ldl, crp), data = cardio)
#> Warning: `outliers` is a screening workflow, not a single hypothesis test.
#> Statistical test workflow
#>
#> Outcome: sbp_3m, ldl, crp
#> Design: outlier screening
#>
#> 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
#> * Effect size not reported.
```

```
#>
#> 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.