visStatistics

Sabine Schilling

2025-05-20

Contents

1 Overview **2**

2 Decision logic **3**

2.1 Numerical response and categorical predictor 3

2.2 Numerical response and numerical feature 4

2.3 Categorical response and categorical feature 4

3 Numerical response and categorical feature **4**

3.1 Categorical feature with two levels: Welch's t-test and Wilcoxon rank-sum test 5

3.1.1 Welch's t-test (`t.test()`) 5

3.1.2 Wilcoxon rank-sum test (`wilcox.test()`) 5

3.1.3 Test choice and graphical output 6

3.1.4 Examples 6

3.2 Categorical feature with more than two levels 9

3.2.1 Fisher's one-way ANOVA (`aov()`) 9

3.2.2 Welch's heteroscedastic one-way ANOVA (`oneway.test()`) 10

3.2.3 Kruskal–Wallis Test (`kruskal.test()`) 10

3.2.4 Test choice 10

3.2.5 Testing the assumption of normality of residual (`shapiro.test()` and `ad.test()`) . . 10

3.2.6 Testing the assumption of equal variances across groups (`bartlett.test()`) 11

3.2.7 Non-parametric alternative: Kruskal-Wallis test 11

3.2.8 Graphical output of ANOVA and one-way test 12

3.2.9 Examples 12

4 Numerical response and numerical feature **17**

4.1 Simple linear regression (`lm()`) 17

4.1.1 Residual analysis 17

4.1.2 Examples 17

5 Categorical response and categorical feature **22**
5.1 Pearson's residuals and mosaic plots . . . 22
5.2 Pearson's χ^2-test (`chisq.test()`) . . . 22
5.3 Pearson's χ^2 test with Yates' continuity correction . . . 22
5.4 Fisher's exact test (`fisher.test()`) . . . 23
5.5 Test choice and graphical output . . . 23
5.6 Transforming a contingency table to a data frame . . . 24
5.7 Examples . . . 24
5.7.1 Pearson's χ^2-test (“) . . . 24
5.7.2 Pearson's χ^2-test with Yate's continuity correction . . . 25
5.7.3 Fisher's exact test (`fisher.test()`) . . . 27
5.8 Saving the graphical output . . . 28

6 Implemented tests **29**
6.1 Numerical response and categorical feature . . . 29
6.1.1 Normality assumption check . . . 29
6.1.2 Homoscedasticity assumption check . . . 29
6.1.3 Post-hoc tests . . . 29
6.2 Numerical response and numerical feature . . . 29
6.3 Categorical response and categorical feature . . . 29

Bibliography **29**

```
library(visStatistics)
```

1 Overview

`visStatistics` automatically selects and visualises appropriate statistical hypothesis tests between a response and a feature variable in a data frame. The choice of test depends on the `class`, distribution, and sample size of the input variables, as well as the user-defined 'conf.level'. The main function `visstat()` visualises the selected test with appropriate graphs (box plots, bar charts, regression lines with confidence bands, mosaic plots, residual plots, Q-Q plots), annotated with the main test results, including any assumption checks and post-hoc analyses. A minimal function call looks like:

```
visstat(dataframe, varsample = "response", varfactor = "feature")
```

The input `data.frame` must be column-based, and the response `varsample` and feature `varfactor` must be character strings naming columns of the `data.frame`.

This scripted workflow is particularly suited for browser-based interfaces that rely on server-side R applications connected to secure databases, where users have no direct access, or for quick data visualisations, e.g., in statistical consulting projects.

This scripted workflow is particularly well suited for interactive interfaces where users access data only through a graphical front end backed by server-side R sessions, as well as for quick data exploration e.g. in statistical consulting contexts.

The remainder of this vignette is organised as follows:

- Section 2 summarises the decision logic of choosing a statistical test, whilst
- Sections 3 - 5 give background on the implemented tests and visualises the decision logic using examples,
- Section 6 gives an overview of the implemented tests.

2 Decision logic

Throughout the remainder, data of class `"numeric"` or `"integer"` are referred as numerical, while data of class `"factor"` are referred to as categorical. The significance level α, used throughout for hypothesis testing, is defined as `1 - conf.level`, where `conf.level` is a user-controllable argument (defaulting to `0.95`).

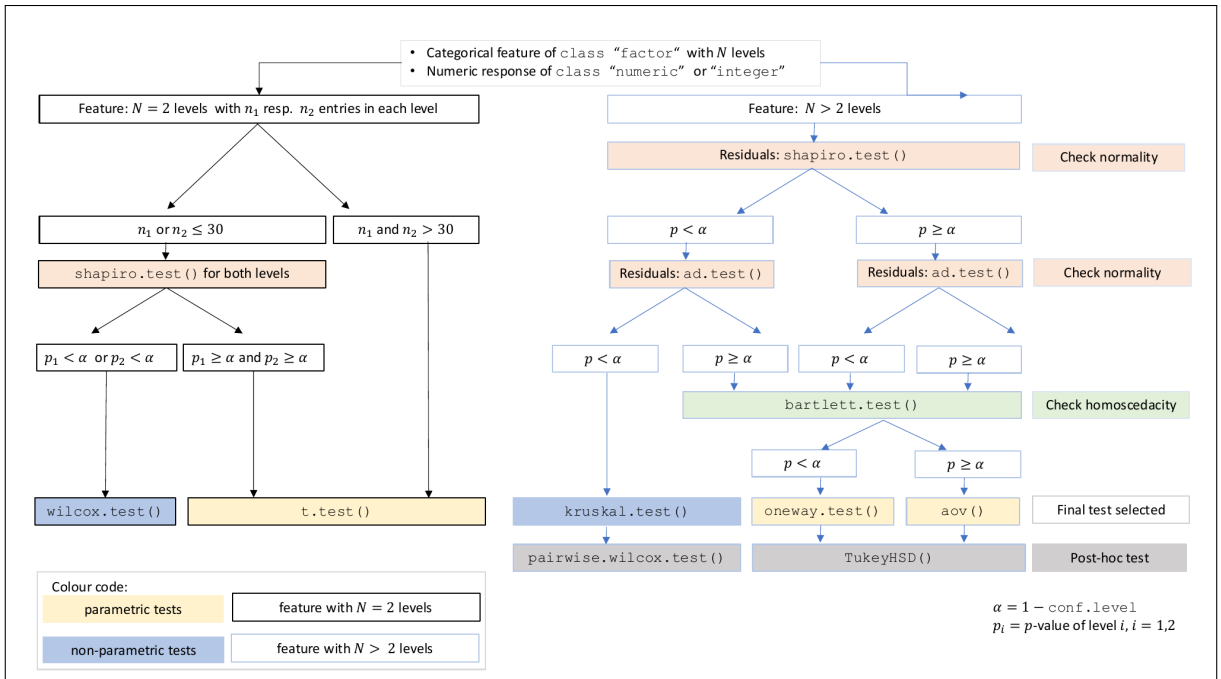
The choice of statistical tests performed by the function `visstat()` depends on whether the data are numerical or categorical, the number of levels in the categorical variable, the distribution of the data, as well as the user-defined 'conf.level'.

The function prioritizes interpretable visual output and tests that remain valid under the following decision logic:

2.1 Numerical response and categorical predictor

When the response is numerical and the predictor is categorical, a statistical hypothesis test of central tendencies is selected.

- If the categorical predictor has exactly two levels, Welch's t-test (`t.test()`) is applied whenever both groups contain more than 30 observations, with the validity of the test supported by the approximate normality of the sampling distribution of the mean under the central limit theorem Lumley et al. (2002). For smaller samples, group - wise normality is assessed using the Shapiro - Wilk test (`shapiro.test()`) at the significance level α. If both groups are found to be approximately normally distributed according to the Shapiro-Wilk test, Welch's t-test is applied; otherwise, the Wilcoxon rank-sum test (`wilcox.test()`) is used.
- For predictors with more than two levels, a model of Fisher's one-way analysis of variables (ANOVA) (`aov()`) is initially fitted. The normality of residuals is evaluated using both the Shapiro-Wilk test (`shapiro.test()`) and the Anderson-Darling test (`ad.test()`); residuals are considered approximately normal if at least one of the two tests yields a result exceeding the significance threshold α. If this condition is met, Bartlett's test (`bartlett.test()`) asseses homoscedasticity. When variances are homogeneous ($p > \alpha$), Fisher's one-way ANOVA (`aov()`) is applied with Tukey's Honestly Significant Differences (HSD) (`TukeyHSD()`) for post-hoc comparison. If variances differ significantly ($p \leq \alpha$), Welch's heteroscedastic one-way ANOVA (`oneway.test()`) is used, also followed by Tukey's HSD. If residuals are not normally distributed according to both tests ($p \leq \alpha$), the Kruskal-Wallis test (`kruskal.test()`) is selected, followed by pairwise Wilcoxon tests (`pairwise.wilcox.test()`). A graphical overview of the decision logic used is provided in the figure below.



Decision tree used to select the appropriate statistical test for a categorical predictor and numerical response, based on the number of factor levels, normality, and homoscedasticity.

Decision tree used to select the appropriate statistical test for a categorical predictor and numerical response, based on the number of factor levels, normality, and homoscedasticity.

2.2 Numerical response and numerical feature

When both the response and predictor are numeric, a simple linear regression model (`lm()`) is fitted and analysed in detail, including residual diagnostics, formal tests, and the plotting of fitted values with confidence bands. Note that **only one** predictor variable is allowed, as the function is designed for two-dimensional visualisation.

2.3 Categorical response and categorical feature

In the case of two categorical variables, `visstat()` tests the null hypothesis that the predictor and response variables are independent using either `chisq.test()` or `fisher.test()`. The choice of test is based on Cochran's rule (Cochran 1954), which advises that the χ^2approximation is reliable only if no expected cell count is less than 1 and no more than 20 percent of cells have expected counts below 5.

Note: Except for the user-adjustable `conf.level` parameter, all statistical tests are applied using their default settings from the corresponding base R functions (e.g., `t.test()`). As a consequence, paired tests are not currently supported. Furthermore, since the main purpose of this package is to visualize statistical test results, only simple linear regression is implemented.

3 Numerical response and categorical feature

If the feature consists of `class` "`factor`" with two or more levels and the response is of `class` "`numeric`" or "`integer`" (both having mode "`numerical`"), statistical tests are applied to compare the central tendencies

across groups. This section describes the conditions under which parametric and non-parametric tests are chosen, based on the response type, the number of factor levels, and the underlying distributional assumptions.

3.1 Categorical feature with two levels: Welch's t-test and Wilcoxon rank-sum test

When the feature variable has exactly two levels, Welch's t-test or the Wilcoxon rank-sum test is applied.

3.1.1 Welch's t-test (`t.test()`)

Welch's t-test assumes that the observations are independent and that the response variable is approximately normally distributed within each group. In contrast to Student's t-test, it does not require the assumption of equal variances (homoscedasticity) between groups. Welch's test remains valid and exhibits only minimal loss of efficiency even when the assumptions of Student's t-test – namely, normality and equal variances of the response variable across groups – are satisfied (Moser and Stevens 1992; Delacre, Lakens, and Leys 2017). Therefore, Student's t-test is not implemented.

Welch's t-test evaluates the null hypothesis that the means of two groups are equal without assuming equal variances. The test statistic is given by (Welch 1947; Satterthwaite 1946)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1^2 and s_2^2 the sample variances, and n_1, n_2 the sample sizes in the two groups. The statistic follows a t-distribution with degrees of freedom approximated by the Welch-Satterthwaite equation:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

The resulting p-value is computed from the t-distribution with ν degrees of freedom.

3.1.2 Wilcoxon rank-sum test (`wilcox.test()`)

The two-sample Wilcoxon rank-sum test (also known as the Mann-Whitney test) (`wilcox.test()`) is a non-parametric alternative that does not require the response variable to be approximately normally distributed within each group. It tests for a difference in location between two independent distributions (Wilcoxon 1945; Mann and Whitney 1947).

The two-level factor variable `varfactor` defines two groups, with sample sizes n_1 and n_2. All $n_1 + n_2$ observations are pooled and assigned ranks from 1 to $n_1 + n_2$. Let W_{Wilcoxon} denote the sum of the ranks assigned to the group corresponding to the first level of `varfactor` containing n_1 observations. The test statistic returned by `visstat()` is then computed as

$$W = W_{\text{Wilcoxon}} - \frac{n_1(n_1 + 1)}{2}.$$

If both groups contain fewer than 50 observations and the data contain no ties, the p-value is computed exactly. Otherwise, a normal approximation with continuity correction is used.

3.1.3 Test choice and graphical output

`visstat` selects between Welch's t-test and the Wilcoxon rank-sum test as follows. If both groups contain more than 30 observations, Welch's t-test (`t.test()`) is always applied, relying on the central limit theorem to justify its application regardless of underlying normality (Rasch, Kubinger, and Moder 2011; Lumley et al. 2002).

If either group contains fewer than 30 observations, the Shapiro-Wilk test (`shapiro.test()`) is applied separately to each group. Welch's t-test is used if both tests do not reject normality at the significance level α; otherwise, the Wilcoxon rank-sum test (`wilcox.test()`) is applied.

The graphical output consists of box plots overlaid with jittered points to display individual observations. When Welch's t-test is applied, the function includes confidence intervals based on the user-specified `conf.level`.

The title is structured as follows:

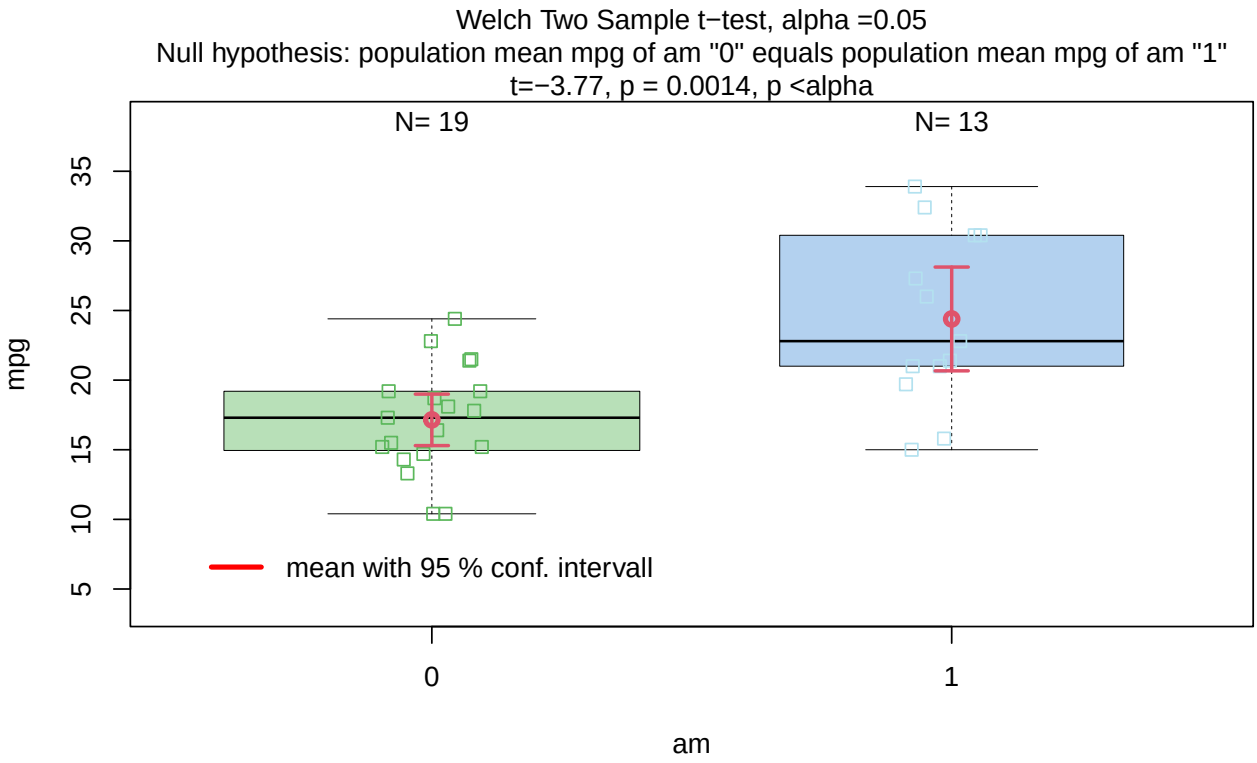
- First line: Test name and chosen significance level α.
- Second line: Null hypotheses automatically adapted based on the user-specified `varsample` and `varfactor`.
- Third line: Test statistic, p-value and automated comparison with α

The function returns a list containing the results of the applied test and the summary statistics used to construct the plot.

3.1.4 Examples

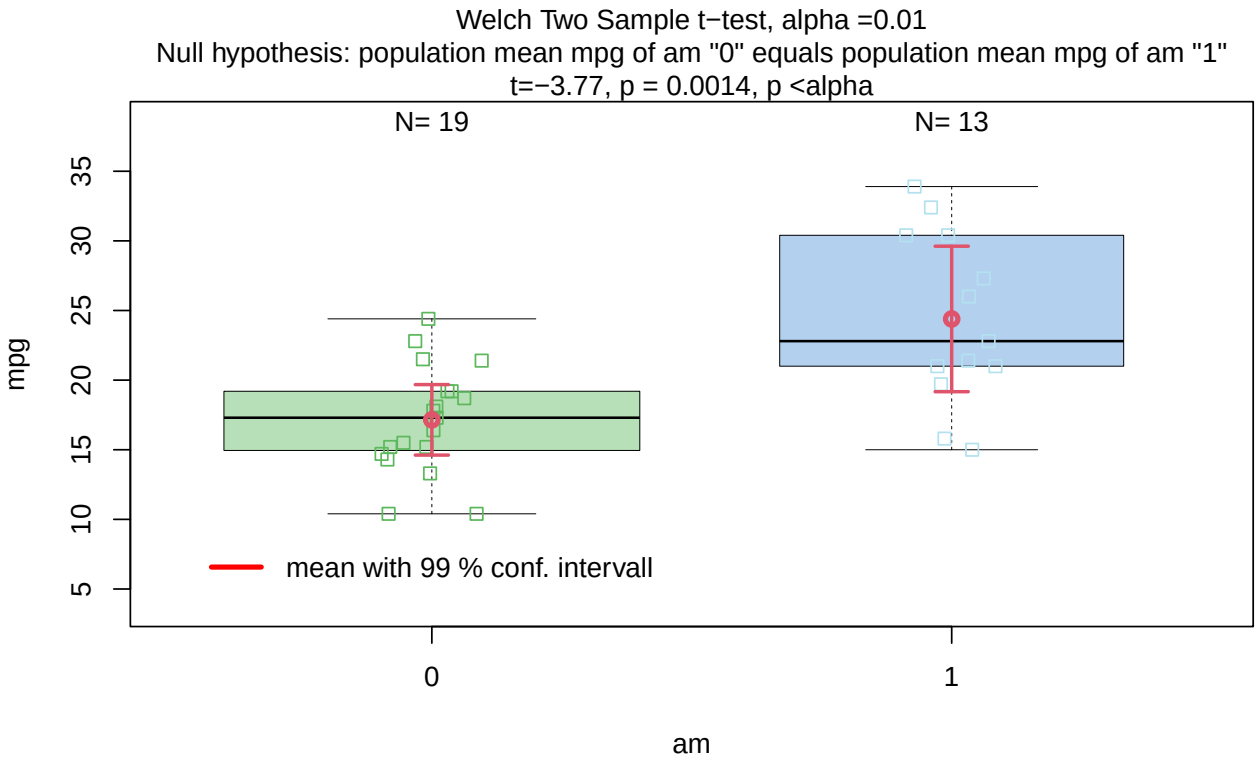
Welch's t-test As an example, we use the *Motor Trend Car Road Tests* dataset (`mtcars`), which contains 32 observations. In the example below, `mpg` denotes miles per (US) gallon, and `am` represents the transmission type (`0` = automatic, `1` = manual).

```
mtcars$am <- as.factor(mtcars$am)
t_test_statistics <- visstat(mtcars, "mpg", "am")
```



Increasing the confidence level `conf.level` from the default 0.95 to 0.99 results in wider confidence intervals, as a higher confidence level requires more conservative bounds to ensure that the interval includes the true parameter value with greater certainty.

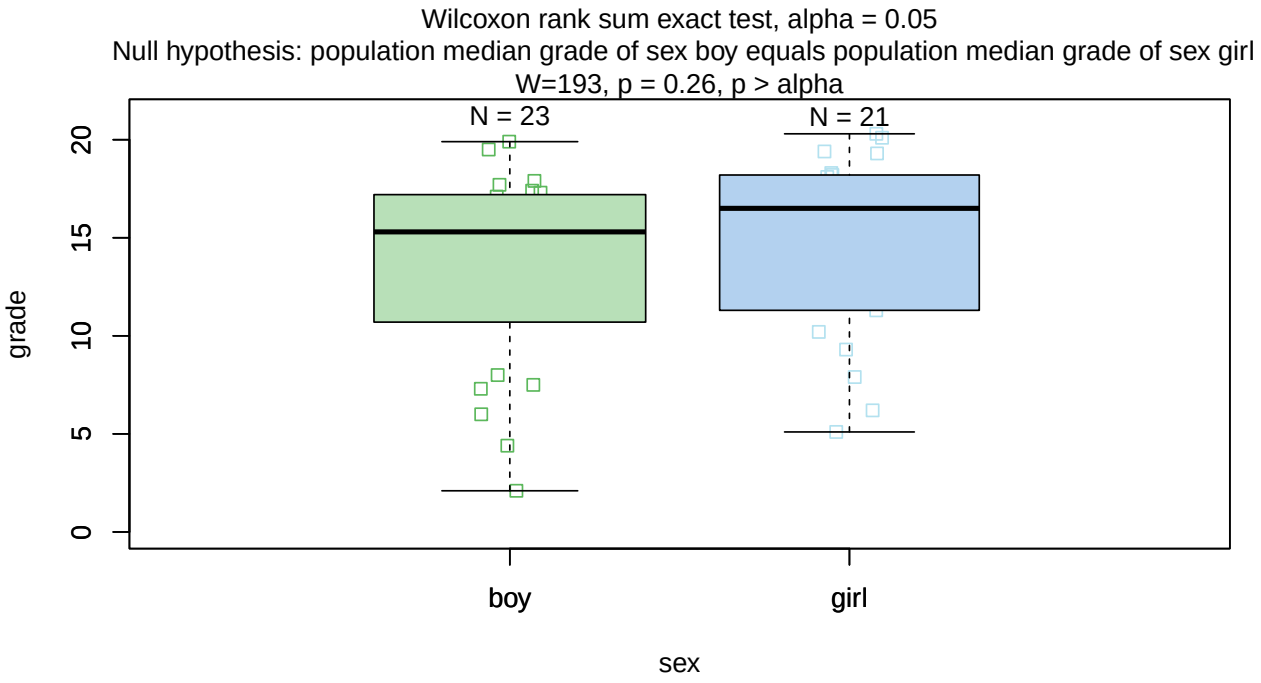
```
mtcars$am <- as.factor(mtcars$am)
t_test_statistics_99 <- visstat(mtcars, "mpg", "am", conf.level = 0.99)
```



Wilcoxon rank sum test The Wilcoxon rank sum test is exemplified on differences between the central tendencies of grades of "boys" and "girls" in a class:

```
grades_gender <- data.frame(
  sex = as.factor(c(rep("girl", 21), rep("boy", 23))),
  grade = c(
    19.3, 18.1, 15.2, 18.3, 7.9, 6.2, 19.4,
    20.3, 9.3, 11.3, 18.2, 17.5, 10.2, 20.1, 13.3, 17.2, 15.1, 16.2, 17.0,
    16.5, 5.1, 15.3, 17.1, 14.8, 15.4, 14.4, 7.5, 15.5, 6.0, 17.4,
    7.3, 14.3, 13.5, 8.0, 19.5, 13.4, 17.9, 17.7, 16.4, 15.6, 17.3, 19.9, 4.4, 2.1
  )
)

wilcoxon_statistics <- visstat(grades_gender, "grade", "sex")
```



3.2 Categorical feature with more than two levels

If the feature is of `class` "`factor`" with **more than two levels** and the response is of `mode` "`numerical`", `visstat()` either performs Fisher's one-way ANOVA (Fisher 1935) (`aov()`), Welch's heteroscedastic one-way ANOVA (Welch 1951) (`oneway.test()`) or, as a non-parametric alternative, the Kruskal-Wallis test (Kruskal and Wallis 1952) (`kruskal.test()`).

In the remainder of this section we will briefly introduce these tests, explain the test choice, demonstrate how the assumptions are tested and illustrate each test by an example.

3.2.1 Fisher's one-way ANOVA (`aov()`)

Fisher's one-way ANOVA (`aov()`) tests the null hypothesis that the means of multiple groups are equal. It assumes independent observations, normally distributed residuals, and **homogeneous** variances across groups. The test statistic is the ratio of the variance explained by differences among group means (between-group variance) to the unexplained variance within groups:

$$F = \frac{\text{between-group variance}}{\text{within-group variance}} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{N-k}},$$

where $\bar{x}_i$ is the mean of group i, $\bar{x}$ is the overall mean, x_{ij} is the observation j in group i, n_i is the sample size in group i, k is the number of groups, and N is the total number of observations.

Under the null hypothesis, this statistic follows an F-distribution with two parameters for degrees of freedom: the numerator $(k-1)$ and the denominator $(N-k)$. The resulting p-value is computed from this distribution.

3.2.2 Welch's heteroscedastic one-way ANOVA (`oneway.test()`)

When the assumption of homoscedasticity is not met, but the assumptions of independent observations and normally distributed residuals are met, Welch's heteroscedastic one-way ANOVA (`oneway.test()`) (Welch 1951) provides an alternative to 'aov()'. It compares group means using weights based on sample sizes and variances. The degrees of freedom are adjusted using a Satterthwaite-type approximation (Satterthwaite 1946), resulting in an F-statistic with non-integer degrees of freedom.

3.2.3 Kruskal–Wallis Test (`kruskal.test()`)

When the assumptions of normality is not met,
`kruskal.test()` provides a non-parametric alternative. It compares group distributions based on ranked values and tests whether the groups come
from the same population (Kruskal and Wallis 1952). The test statistic is defined as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2,$$

where n_i is the sample size in group i, k is the number of groups, $\bar{R}_i$ is the average rank of group i, N is the total sample size, and $\bar{R} = \frac{N+1}{2}$ is
the average of all ranks. Under the null hypothesis, H approximately follows
a χ^2 distribution with $k-1$ degrees of freedom.

3.2.4 Test choice

The test logic follows directly from above assumptions for `aov()` and `oneway.test()`: `visstat()` initially attempts an analysis of variance (ANOVA) as implemented in `aov()`. It is performed only if **both** of the following null hypotheses cannot be rejected at the specified `conf.level`:

1. The standardized residuals follow a normal distribution, and
2. The residuals exhibit homoscedasticity (equal variances across groups).

If only the assumption of normality is satisfied, `visstat()` applies Welch's one-way test (`oneway.test()`).

If the normality assumption is violated, a Kruskal-Wallis test (`kruskal.test()`) is used instead.

These assumptions are tested internally using the `visAnovaAssumptions()` function.

3.2.5 Testing the assumption of normality of residual (`shapiro.test()` and `ad.test()`)

The `visAnovaAssumptions()` function assesses the normality of standardized residuals from the ANOVA fit using both the Shapiro-Wilk test (`shapiro.test()`) and the Anderson-Darling test (`ad.test()`). Normality is assumed if at least one of the two tests yields a p-value greater than α.

The function generates two diagnostic plots: - a scatterplot of the standardized residuals against the fitted means of the linear model for each level of the feature (`varfactor`), and
- a Q-Q plot of the standardized residuals.

3.2.6 Testing the assumption of equal variances across groups (`bartlett.test()`)

Both `aov()` and `oneway.test()` assess whether two or more samples drawn from normal distributions have the same mean. While `aov()` assumes homogeneity of variances across groups, `oneway.test()` does not require the variances to be equal.

Homoscedasticity is assessed using Bartlett's test (`bartlett.test()`\), which tests the null hypothesis that the variances across all levels of the grouping variable are equal.

Post-hoc analysis: Tukey's Honestly Significant Differences (HSD) and Sidak-corrected confidence intervals Simple pairwise comparisons of group means following an analysis of variance increase the probability of incorrectly declaring a significant difference when, in fact, there is none.

This inflation of error is quantified by the family-wise error rate, which refers to the probability of making at least one Type I error, that is, falsely rejecting the null hypothesis across all pairwise comparisons.

To control it, `visstat()` performs post-hoc analysis using Tukey's Honestly Significant Difference (HSD) test and displays Sidak-corrected confidence intervals.

The `visstat()` function controls the probability of a Type I error by applying Tukey's Honestly Significant Differences procedure, as implemented in `TukeyHSD()`. Based on the specified confidence level (`conf.level`), it constructs a set of confidence intervals for all pairwise differences between factor level means. A significant difference between two means is indicated when the corresponding confidence interval does not include zero. The function returns both the HSD-adjusted p-values and the associated confidence intervals for all pairwise comparisons.

In the graphical output for the one-way test and ANOVA, green letters displayed below each group summarize the results of the Tukey HSD post-hoc test: Two groups are considered significantly different if they are assigned different letters, indicating a Tukey HSD-adjusted p-value smaller than α.

Tukey's HSD procedure is based on pairwise comparisons of the differences between the means at each factor level and produces a set of corresponding confidence intervals. The Sidak procedure, on the other hand, addresses the problems of a type I error by lowering the acceptable probability of a type I error for all comparisons of the levels of the independent, categorical variable.

The Sidak-corrected acceptable probability of error (Šidák 1967) is defined as $\alpha_{Sidak} = (1\text{-}\texttt{conf.int})^{1/M}$, where $M = \frac{n\cdot(n-1)}{2}$ is the number of pairwise comparisons of the n levels of the categorical variable.

In the graphical display of One-way test and ANOVA, `visstat()` displays both the `conf.level` $\cdot$ 100%-confidence intervals alongside the larger, Sidak-corrected $(1-\alpha_{Sidak})\cdot 100$ %-confidence intervals.

Note that the current structure of `visstat()` does not allow the study of interactions between the different levels of an independent variable.

3.2.7 Non-parametric alternative: Kruskal-Wallis test

If the p-value from the Shapiro-Wilk test (`shapiro.test()`) applied to the standardized residuals is smaller than the significance level α, `visstat()` selects a non-parametric alternative: the Kruskal-Wallis rank sum test.

The function `kruskal.test()` tests the null hypothesis that the group medians are equal across all levels of the categorical feature.

It compares group distributions based on ranked values and tests whether the groups come from the same population

Post-hoc analysis: `pairwise.wilcox.test()` As a post-hoc analysis following the Kruskal-Wallis test, `visstat()` applies the pairwise Wilcoxon rank sum test using `pairwise.wilcox.test()`, with Holm's method as the default adjustment for multiple comparisons (Holm 1979).

If the Holm-adjusted p-value for a given pair of groups is smaller than the significance level α, the green letters displayed below the corresponding box plots will differ. Otherwise, the groups are considered not significantly different.

Apart from the multiple comparison adjustment, the graphical representation of the Kruskal-Wallis result is similar to that used for the Wilcoxon rank sum test.

The function returns a list containing the test statistic from the Kruskal-Wallis rank sum test, along with the Holm-adjusted p-values for all pairwise comparisons.

3.2.8 Graphical output of ANOVA and one-way test

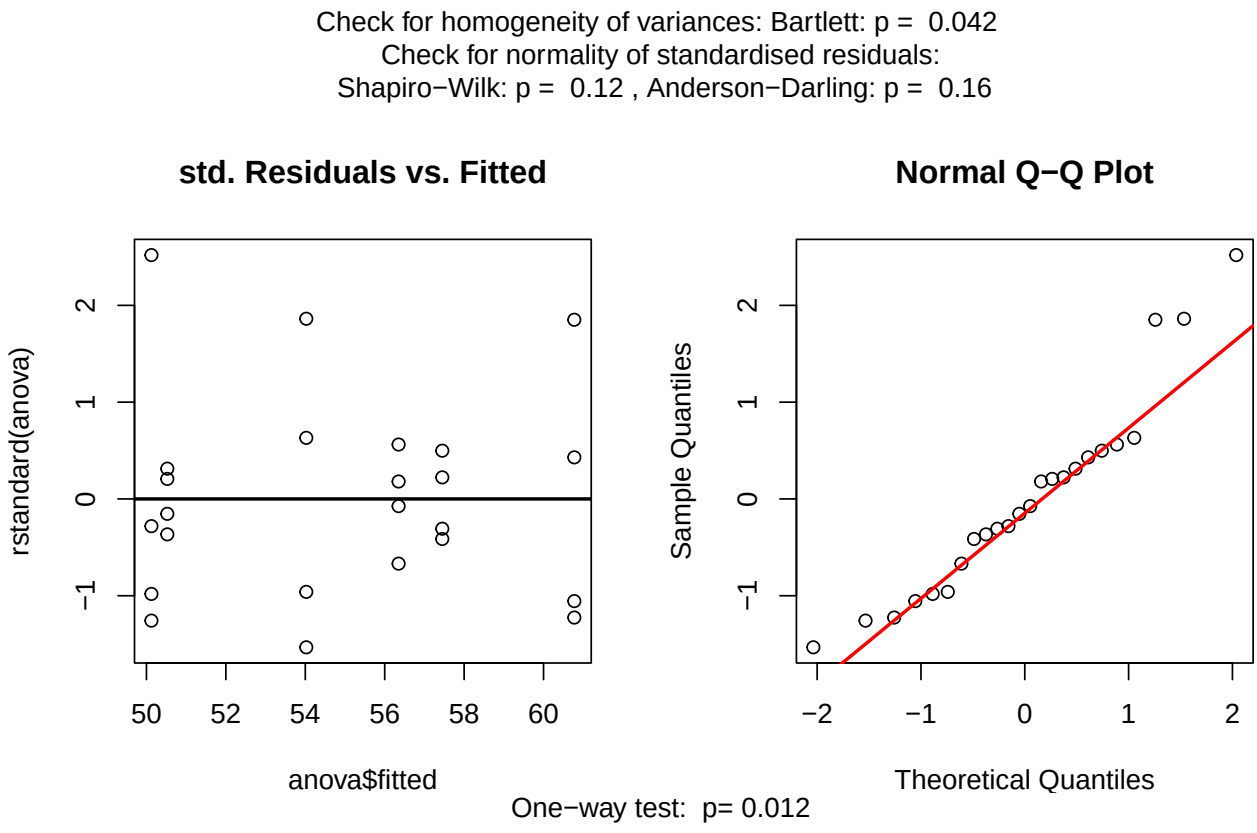
Depending on the p-value returned by `bartlett.test()`, the corresponding test is selected and its p-value is displayed in the figure title:

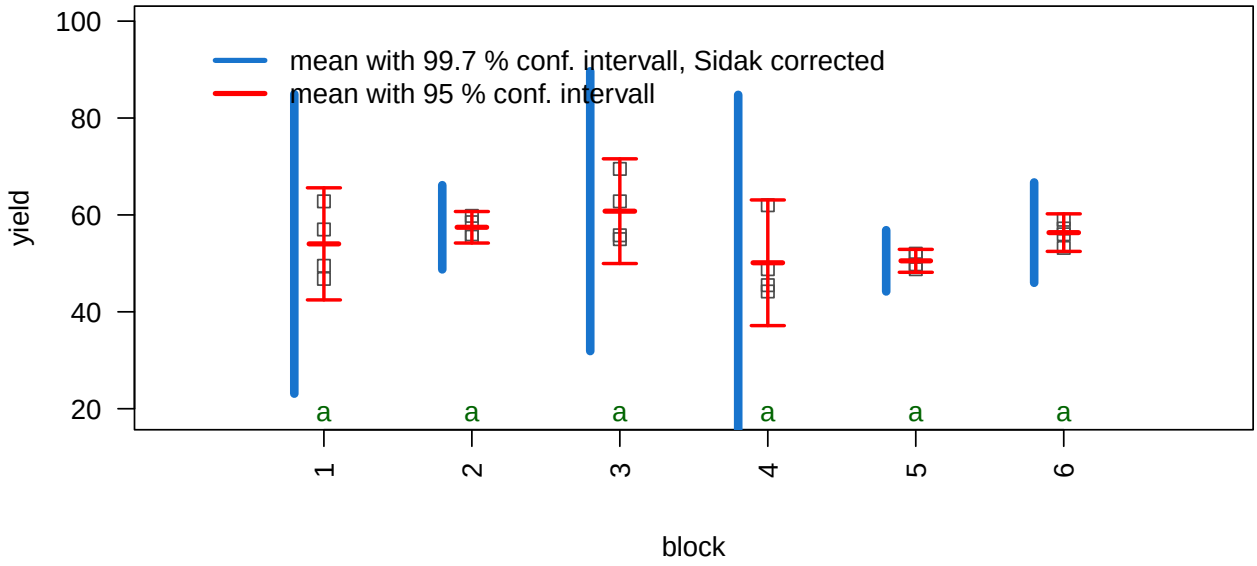
- If the p-value of `bartlett.test()` is greater than α, we assume homogeneity of variances across groups and report the p-value from the analysis of variance (ANOVA) implemented as (`aov()`).
- Otherwise, homoscedasticity cannot be assumed, and the function reports the p-value from Welch's one-way test (`oneway.test()`).

3.2.9 Examples

One-way test The `npk` dataset reports the yield of peas (in pounds per block) from an agricultural experiment conducted on six blocks. In this experiment, the application of three different fertilisers – nitrogen (N), phosphate (P), and potassium (K) – was varied systematically. Each block received either none, one, two, or all three of the fertilisers,

```
oneway_npk <- visstat(npk, "yield", "block")
```

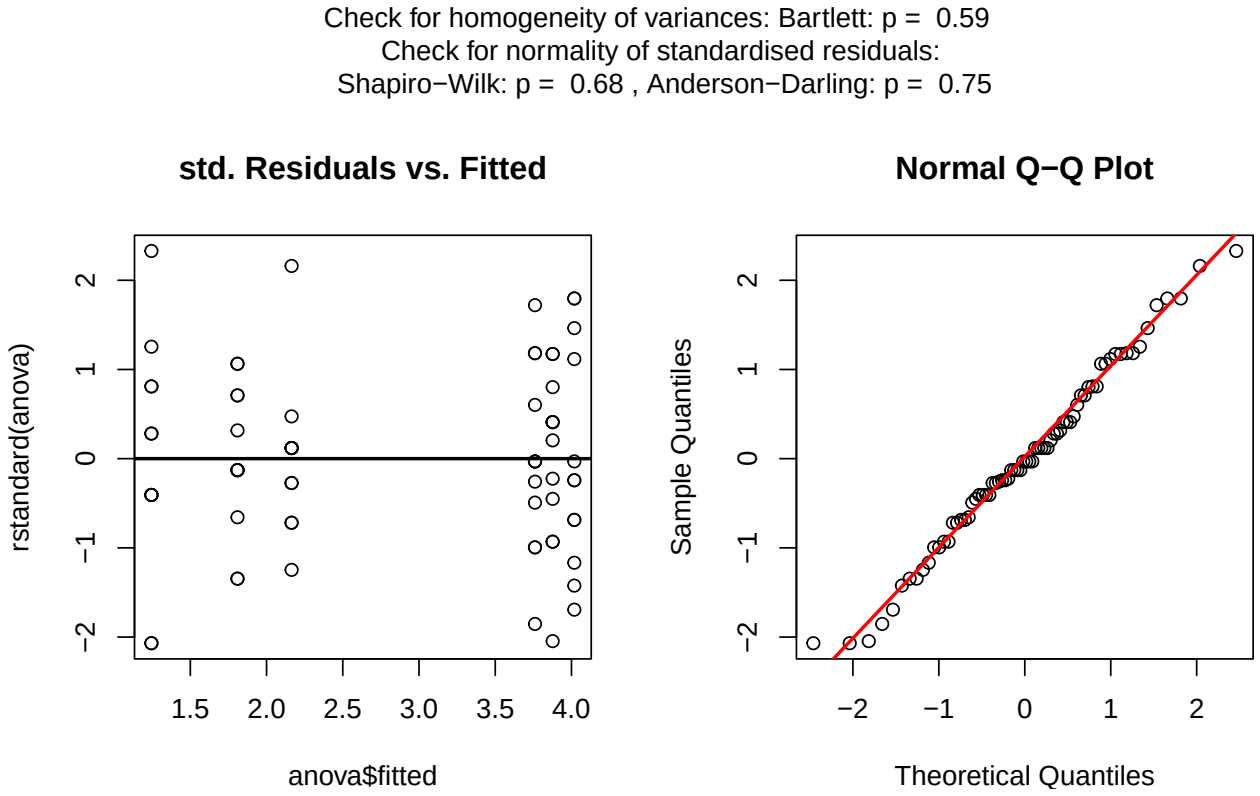


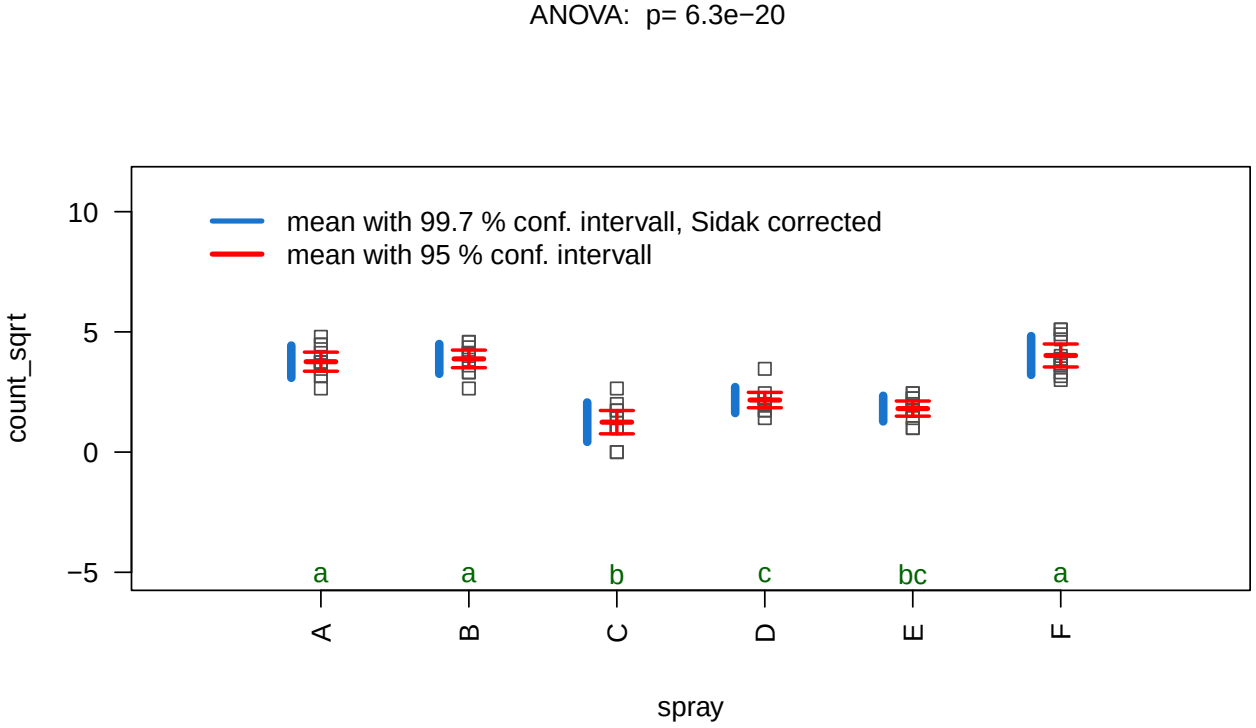


Normality of residuals is supported by graphical diagnostics (scatter plot of standardized residuals, Q-Q plot) and formal tests (Shapiro-Wilk and Anderson-Darling, both with $p > \alpha$). However, homogeneity of variances is not supported at the given confidence level ($p < \alpha$, `bartlett.test()`), so the p-value from the variance-robust `oneway.test()` is reported. Post-hoc analysis with `TukeyHSD()` shows no significant yield differences between blocks, as all share the same group label (e.g., all green letters).

ANOVA example The `InsectSprays` dataset reports insect counts from agricultural experimental units treated with six different insecticides. To stabilise the variance in counts, we apply a square root transformation to the response variable.

```
insect_sprays_tr <- InsectSprays
insect_sprays_tr$count_sqrt <- sqrt(InsectSprays$count)
visstat(insect_sprays_tr, "count_sqrt", "spray")
```

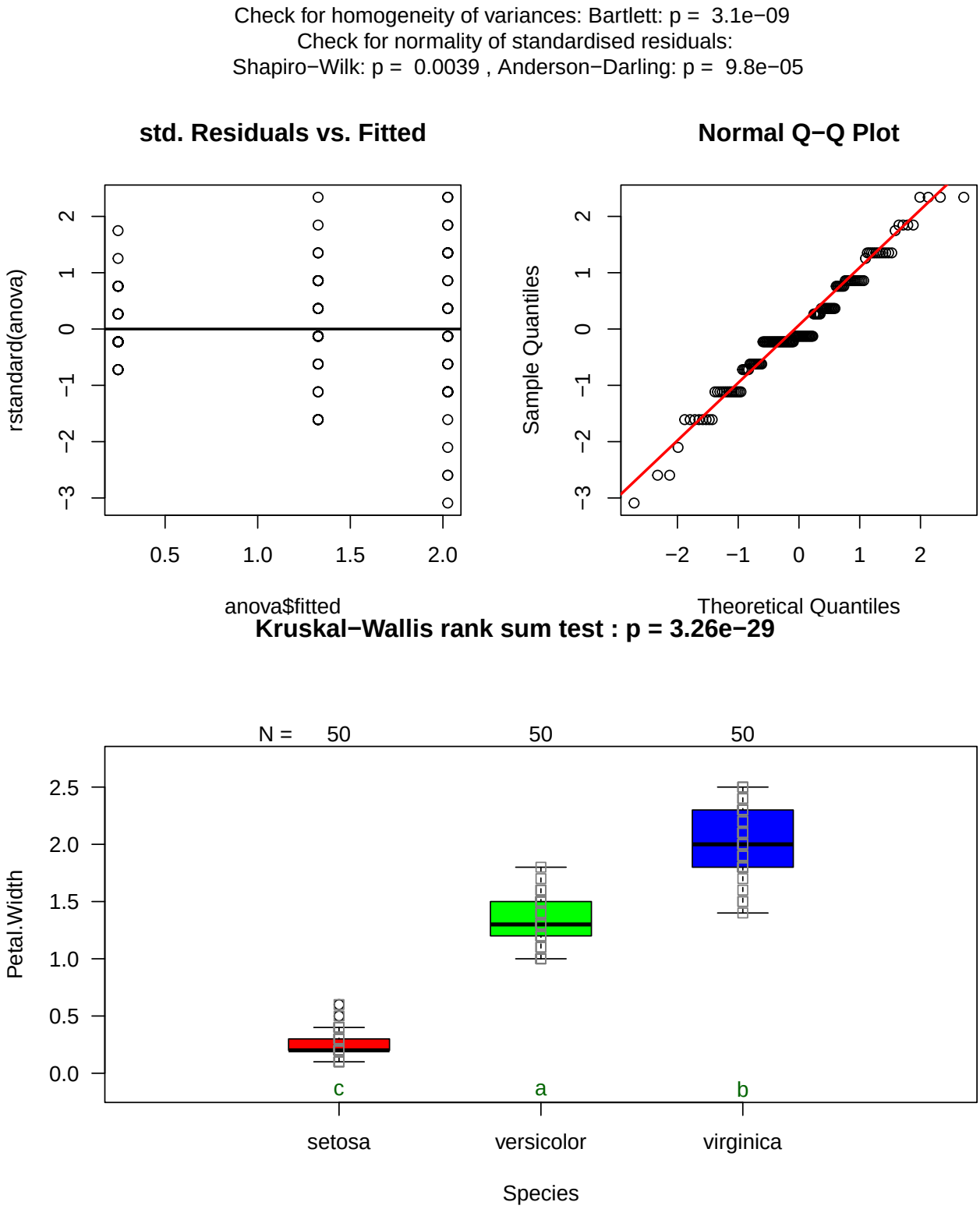




After the transformation, the homogeneity of variances can be assumed ($p > \alpha$ as calculated with the `bartlett.test()`), and the p-value of the `aov()` is displayed.

Kruskal-Wallis rank sum test The `iris` dataset contains petal width measurements (in cm) for three different iris species.

```
visstat(iris, "Petal.Width", "Species")
```



In this example, scatter plots of the standardised residuals and the Q-Q plot suggest that the residuals are not normally distributed. This is confirmed by very small p-values from both the Shapiro-Wilk and Anderson-Darling tests.

If both p-values are below the significance level α, `visstat()` switches to the non-parametric `kruskal.test()`. Post-hoc analysis using `pairwise.wilcox.test()` shows significant differences in

petal width between all three species, as indicated by distinct group labels (all green letters differ).

4 Numerical response and numerical feature

4.1 Simple linear regression (`lm()`)

If the feature `varfactor` and the response `varsample` are both numerical and *contain only one level each* , `visstat()` performs a simple linear regression.

The resulting regression plot displays the point estimation of the regression line

$y = a + b \cdot x$, where where y is the response variable, x is the feature variable, a the intercept and b the slope pf the regression line

Note that multiple linear regression is not implemented, as the main focus of `visstat()` is on the visualisation of statistical tests.

4.1.1 Residual analysis

`visstat()` checks the normality of the standardised residuals from `lm()` both graphically and using the Shapiro-Wilk and Anderson-Darling tests. If the p-values for the null hypothesis of normally distributed residuals from both tests are smaller than $1-$`conf.int`, the title of the residual plot will display the message: "Requirement of normally distributed residuals not met".

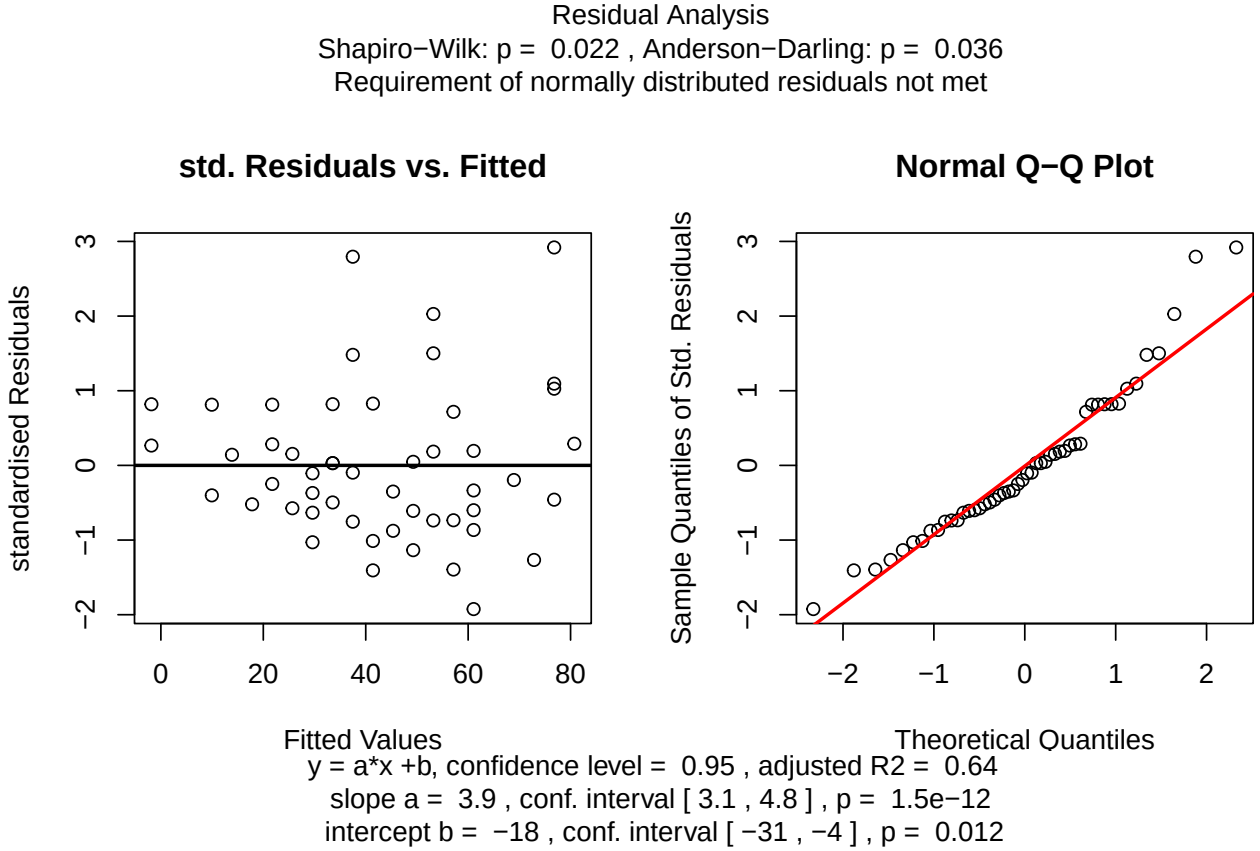
Regardless of the result of the residual analysis, `visstat()` proceeds to perform the regression. The title of the graphical output indicates the chosen confidence level (`conf.level`), the estimated regression parameters with their confidence intervals and p-values, and the adjusted R^2. The plot displays the raw data, the fitted regression line, and both the confidence and prediction bands corresponding to the specified `conf.level`.

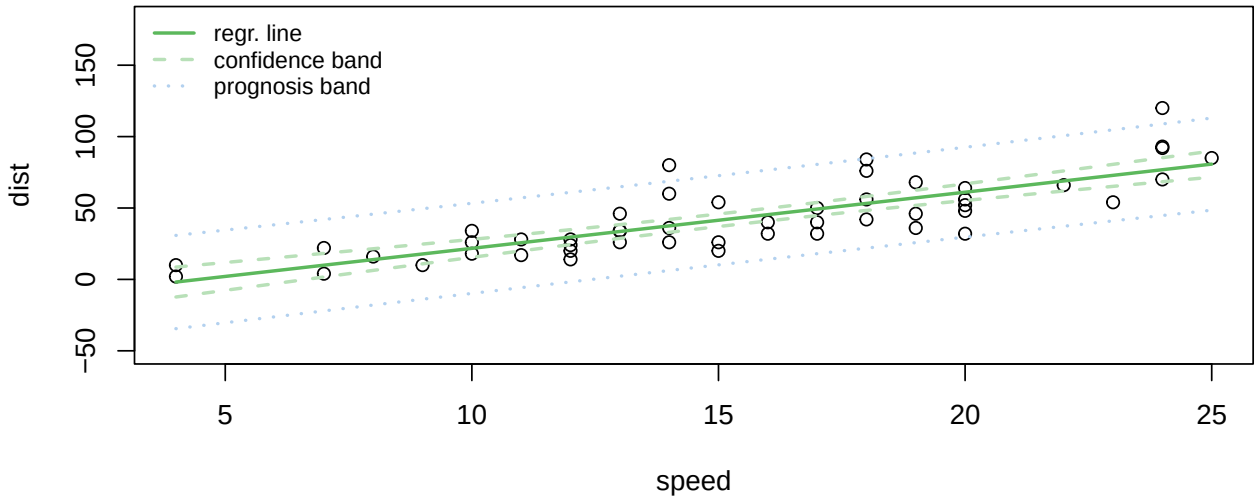
`visstat()` returns a list containing the regression test statistics, the p-values from the normality tests of the standardised residuals, and the pointwise estimates of the confidence and prediction bands.

4.1.2 Examples

4.1.2.1 dataset: `cars` The `cars` dataset reports the speed of cars in miles per hour (`speed`) and the stopping distance in feet (`dist`).

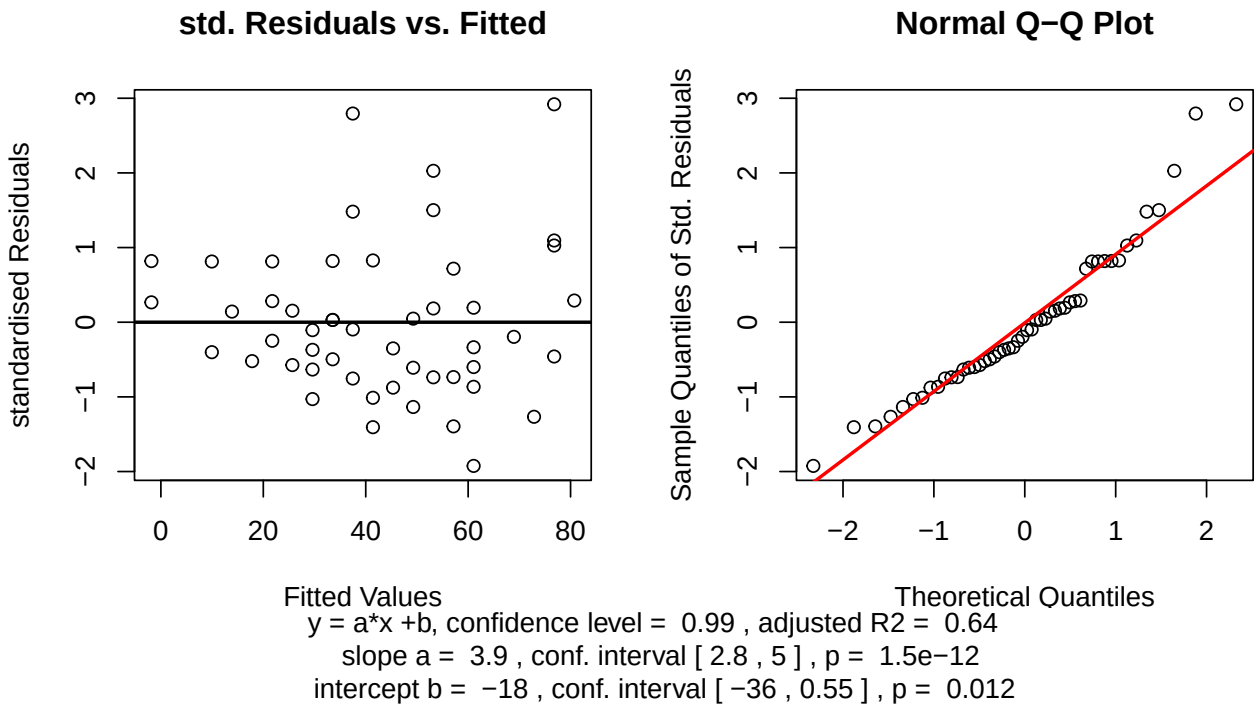
```
linreg_cars <- visstat(cars, "dist", "speed")
```

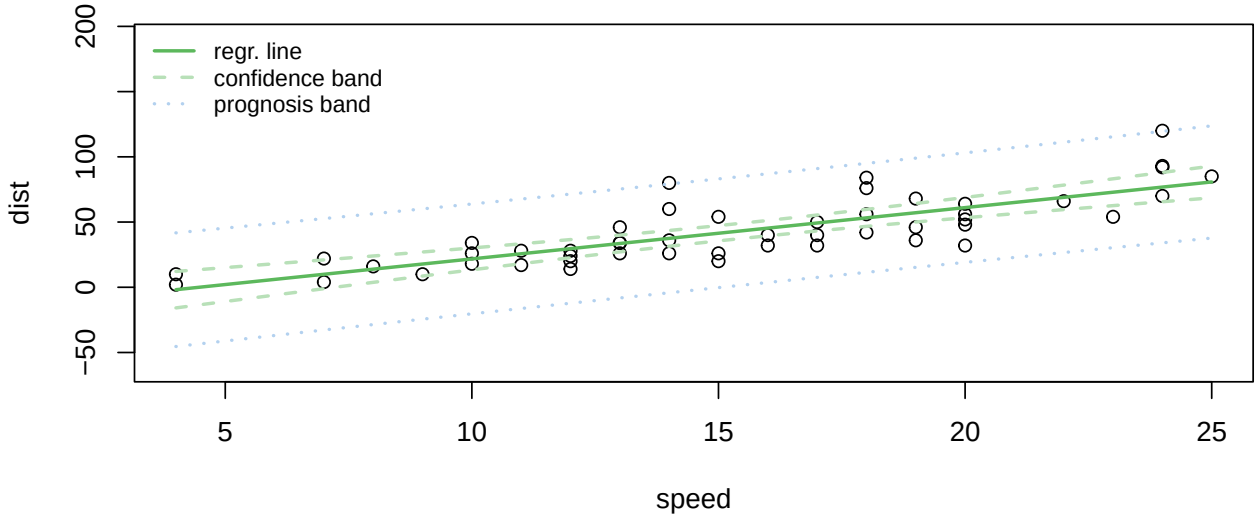




Increasing the confidence level `conf.level` from the default 0.95 to 0.99 results in wider confidence and prediction bands:

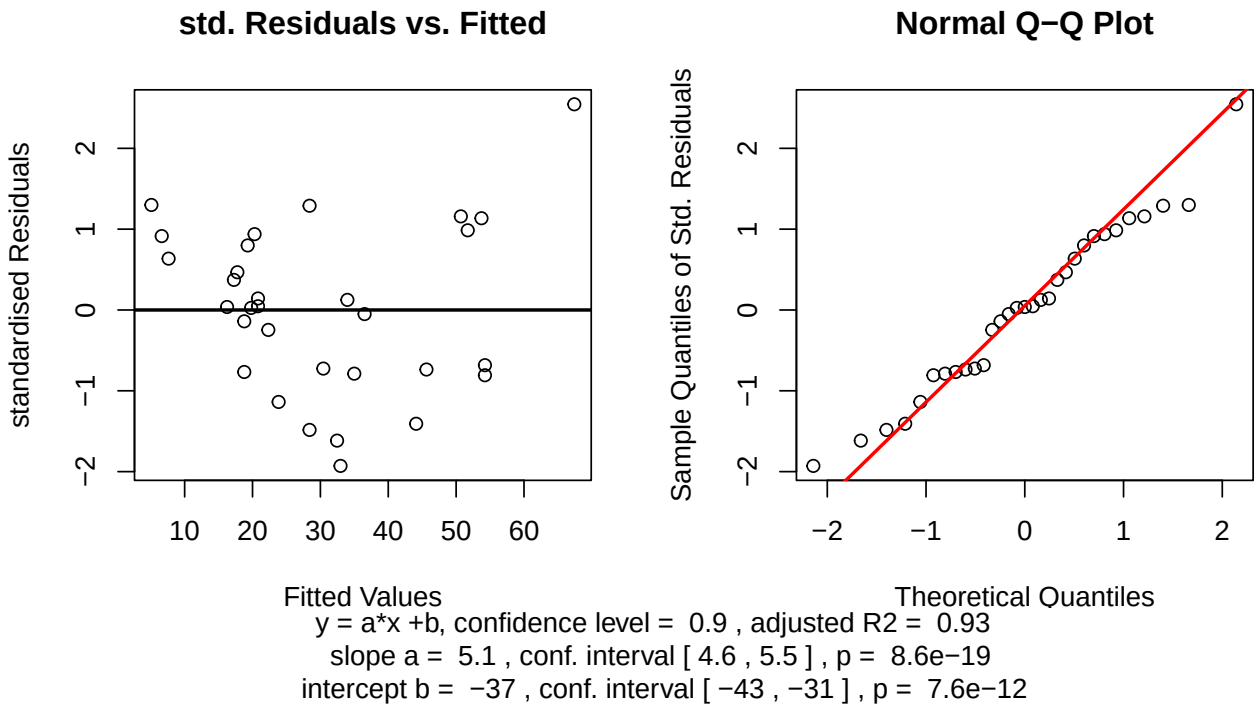
```
linreg_cars <- visstat(cars, "dist", "speed", conf.level = 0.99)
```

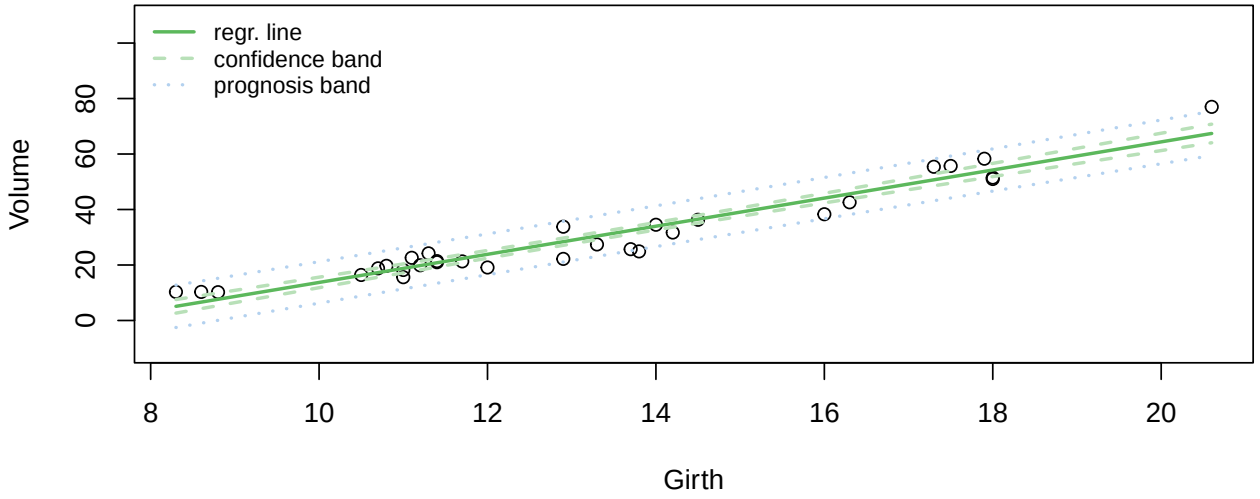




p-values greater than `conf.level` in both the Anderson-Darling normality test and the Shapiro-Wilk test of the standardised residuals indicate that the normality assumption of the residuals underlying the linear regression is met.

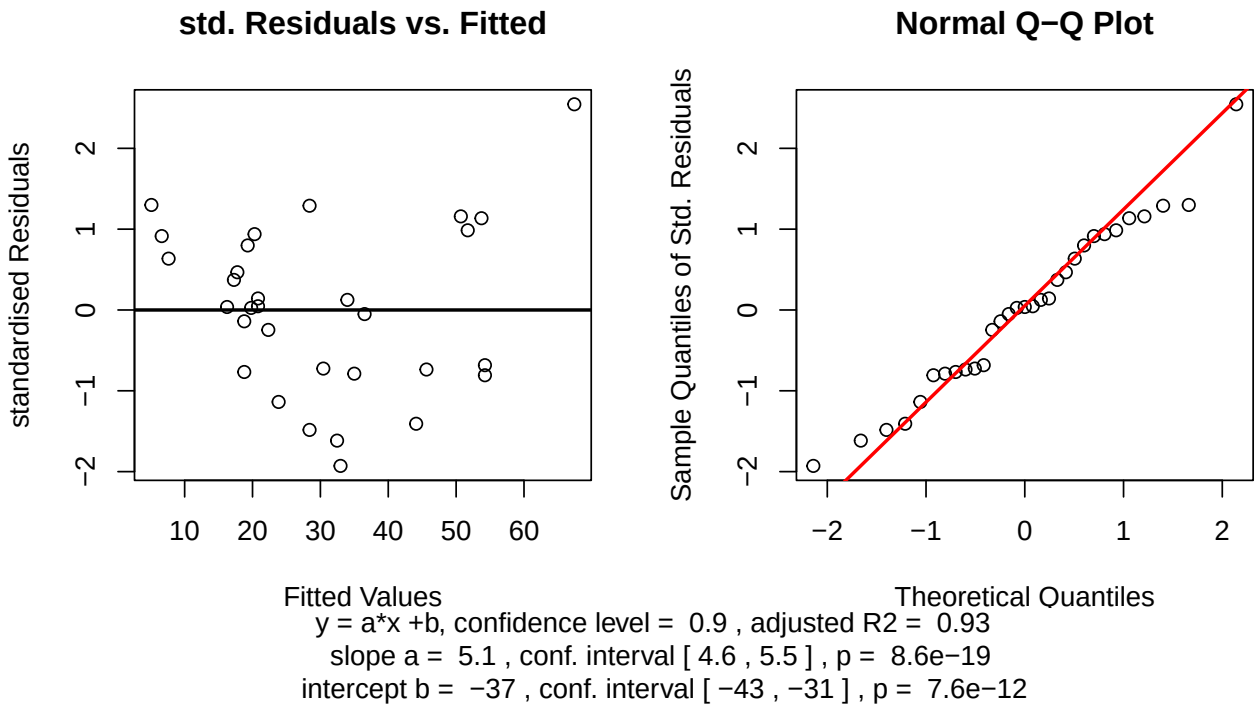
```
linreg_trees <- visstat(trees, "Volume", "Girth", conf.level = 0.9)
```

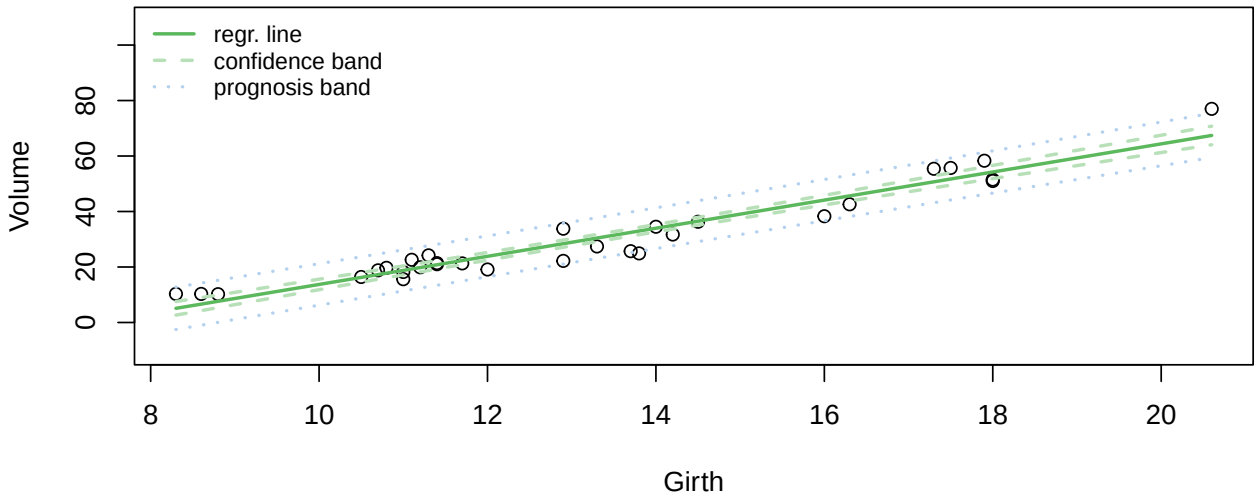




4.1.2.2 dataset: `trees` The `trees` dataset provides measurements of the diameter (`Girth`, in inches), `Height` (in feet), and `Volume` (in cubic feet) of black cherry trees. In this example, we choose `Volume` as the response and `Girth` as the feature.

```
linreg_cars <- visstat(trees, "Volume", "Girth", conf.level = 0.9)
```





Both the graphical analysis of the standardised residuals and p-values greater than α in the Anderson-Darling and Shapiro-Wilk tests suggest that the assumption of normally distributed residuals is met. Furthermore, the linear regression model explains 93% of the total variance in the response variable `Volume`.

5 Categorical response and categorical feature

When both `varfactor` and `varsample` are categorical (i.e., of class `factor`), `visstat()` tests the null hypothesis that the two variables are independent. Observed frequencies are typically arranged in a contingency table, where rows index the levels i of the response variable and columns index the levels j of the feature variable.

5.1 Pearson's residuals and mosaic plots

Mosaic plots provide a graphical representation of contingency tables, where the area of each tile is proportional to the observed cell frequency. To aid interpretation, tiles are coloured based on Pearson residuals from a chi-squared test of independence. These residuals measure the standardised deviation of observed from expected counts under the null hypothesis of independence.

Let O_{ij} and E_{ij} denote the observed and expected frequencies in row i and column j of an $R \times C$ contingency table. The Pearson residual for each cell is defined as

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}, \quad i = 1, \ldots, R, \quad j = 1, \ldots, C.$$

Positive residuals (shaded in blue) indicate observed counts greater than expected, while negative values suggest under-representation (shaded in red). Colour shading thus highlights which combinations of categorical levels contribute most to the overall association.

5.2 Pearson's χ^2-test (`chisq.test()`)

The test statistic of Pearson's χ^2-test (Pearson 1900) is the sum of squared Pearson residuals:

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} r_{ij}^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The test statistic is compared to the chi-squared distribution with $ (R - 1)(C - 1)$ degrees of freedom. The resulting p-value corresponds to the upper tail probability — that is, the probability of observing a value greater than or equal to the test statistic under the null hypothesis.

5.3 Pearson's χ^2 test with Yates' continuity correction

Yates' correction is applied to the Pearson χ^2 statistic in 2×2 contingency tables (with one degree of freedom). In this case, the approximation of the discrete sampling distribution by the continuous χ^2 distribution tends to overestimate the significance level of the test. To correct for this, Yates proposed subtracting 0.5 from each absolute difference between observed and expected counts (Yates 1934), resulting in a smaller test statistic:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

This reduced test statistic yields a larger p-value, thereby lowering the risk of a type I error.

5.4 Fisher's exact test (`fisher.test()`)

The χ^2 approximation is considered reliable only if no expected cell count is less than 1 and no more than 20 percent of cells have expected counts below 5 (Cochran 1954)). If this condition is not met, Fisher's exact test (Fisher 1935) (`fisher.test()`) is applied instead, as it is a non-parametric method that does not rely on large-sample approximations. The test calculates an exact
p-value for testing independence by conditioning on the observed margins:
the row totals $R_i = \sum_{j=1}^{C} O_{ij}$ and the column totals
$C_j = \sum_{i=1}^{R} O_{ij}$, defining the structure of the
contingency table.

In the 2×2 case, the observed table can be written as:

	C_1	C_2	Row sums
R_1	a	b	$a+b$
R_2	c	d	$c+d$
Column sums	$a+c$	$b+d$	n

Let

$$O = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

denote the above observed 2×2 contingency table.
The exact probability of observing this table under the null hypothesis of
independence, given the fixed margins, is given by the hypergeometric probability mass function (PMF)

$$\mathbb{P}(O \mid R_1, R_2, C_1, C_2) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}},$$

where $n = a + b + c + d$ is the total sample size.

The p-value is computed by summing the probabilities of all tables
with the same margins whose probabilities under the null are less than
or equal to that of the observed table.

For general $R \times C$ tables, `fisher.test()` generalises this
approach using the multivariate hypergeometric distribution.

5.5 Test choice and graphical output

If the expected frequencies are sufficiently large - specifically, if at least 80% of the cells have expected counts greater than 5 and no expected count is zero - the function uses Pearson's χ^2-test (`chisq.test()`).

Otherwise, it switches to Fisher's exact test (`fisher.test()`) (Cochran 1954).

For 2-by-2 contingency tables, Yates' continuity correction (Yates 1934) is always applied to Pearson's χ^2-test.

For all tests of independence `visstat()` displays a grouped column plot that includes the respective test's p-value in the title, as well as a mosaic plot showing colour-coded Pearson residuals and the p-value of Pearson's χ^2-test.

5.6 Transforming a contingency table to a data frame

The following examples for tests of categorical feature and response are all based on the `HairEyeColor` contingency table.

Contingency tables must be converted to the required column-based `data.frame` using the helper function `counts_to_cases()`. The function transforms the contingency table `HairEyeColor` into `data.frame` named `HairEyeColourDataFrame`.

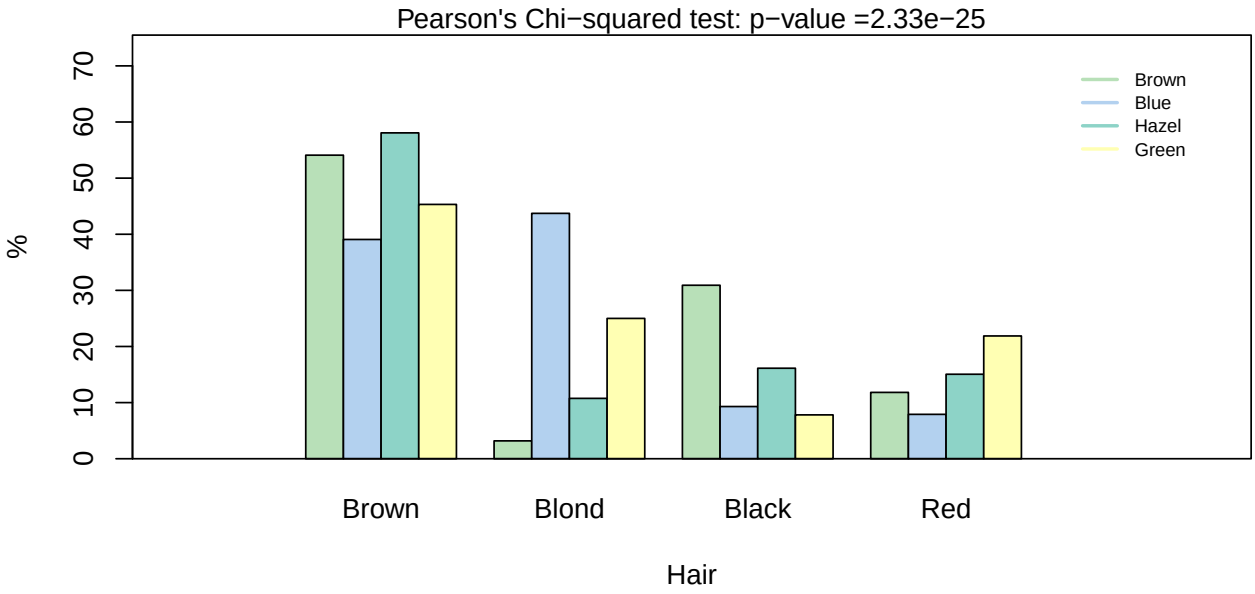
```
HairEyeColourDataFrame <- counts_to_cases(as.data.frame(HairEyeColor))
```

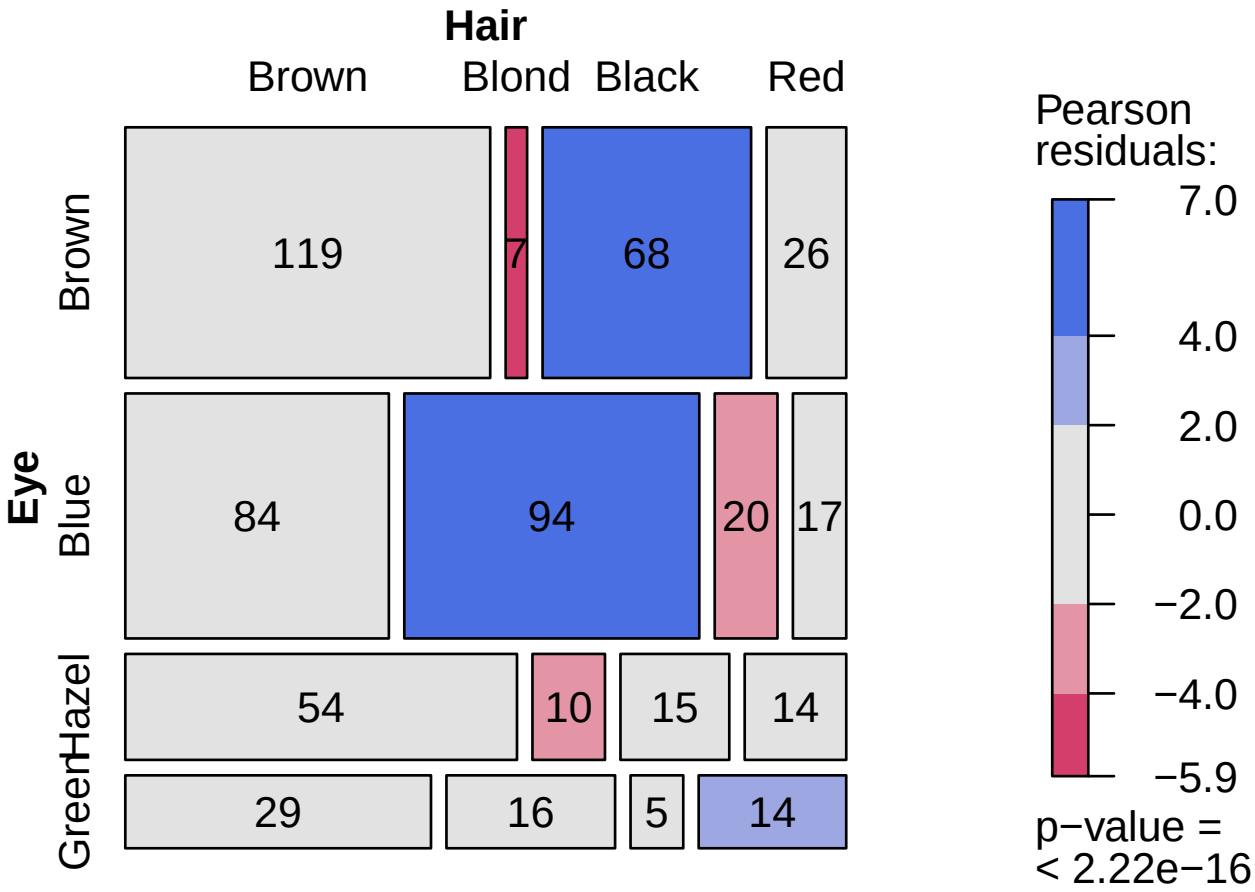
5.7 Examples

In all examples of this section, we will test the null hypothesis that hair colour ("Hair") and eye colour ("Eye") are independent of each other.

5.7.1 Pearson's χ^2-test (")

```
hair_eye_colour_df <- counts_to_cases(as.data.frame(HairEyeColor))
visstat(hair_eye_colour_df, "Hair", "Eye")
```



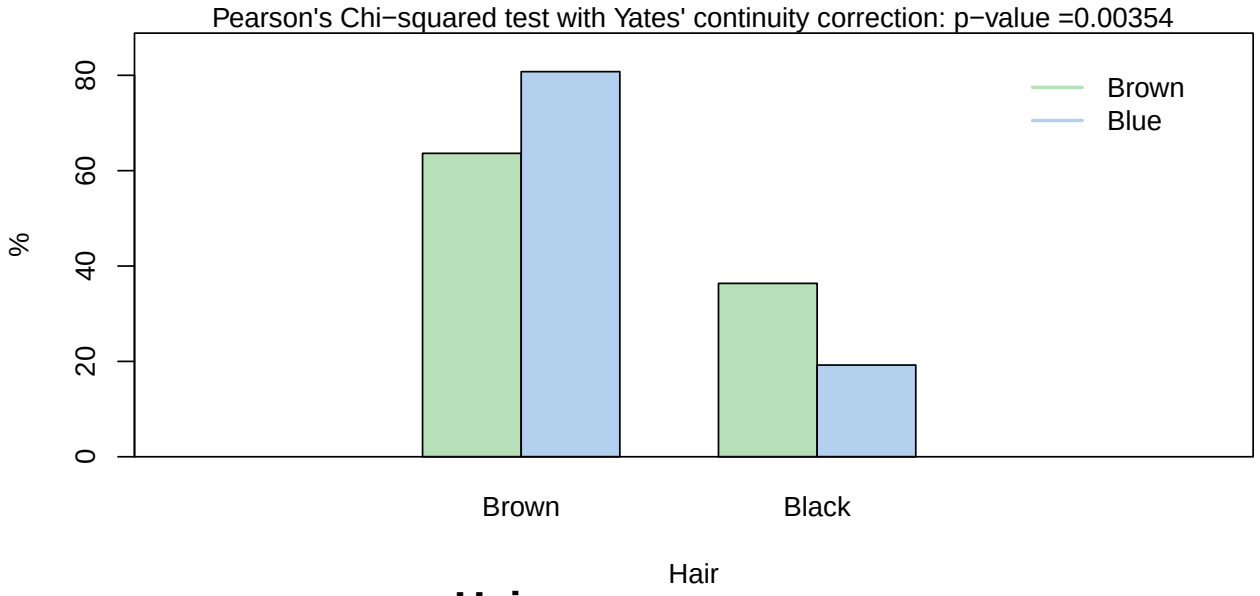


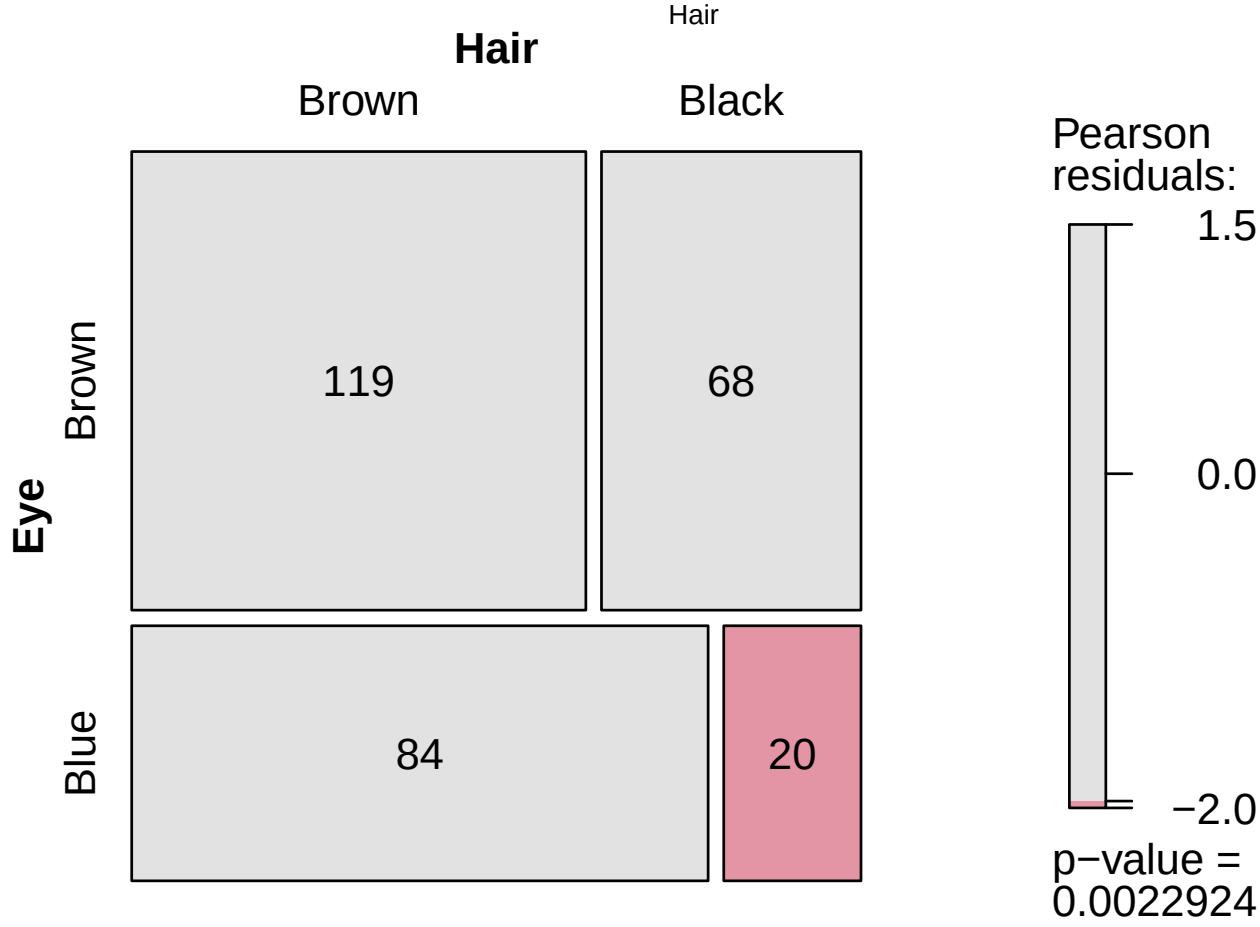
The graphical output shows that the null hypothesis of Pearson's χ^2 test – namely, that hair colour and eye colour are independent – must be rejected at the default significance level $\alpha = 0.05$ ($p = 2.33 \cdot 10^{-25} < \alpha$). The mosaic plot indicates that the strongest deviations are due to over-representation of individuals with black hair and brown eyes, and of those with blond hair and blue eyes. In contrast, individuals with blond hair and brown eyes are the most under-represented.

5.7.2 Pearson's χ^2-test with Yate's continuity correction

In the following example, we restrict the data to participants with either black or brown hair and either brown or blue eyes, resulting in a 2-by-2 contingency table.

```
hair_black_brown_eyes_brown_blue <- HairEyeColor[1:2, 1:2, ]
# Transform to data frame
hair_black_brown_eyes_brown_blue_df <- counts_to_cases(as.data.frame(hair_black_brown_eyes_brown_blue))
# Chi-squared test
visstat(hair_black_brown_eyes_brown_blue_df, "Hair", "Eye")
```





Also in this reduced dataset we reject the null hypothesis of independence of the hair colors "brown" and "black" from the eye colours "brown" and " blue". The mosaic plot shows that blue eyed persons with black hair are under-represented. Note the higher p-value of Pearson's χ^2-test with Yate's continuity correction (p=0.00354) compared to the p-value of Pearson's χ^2-test (p=0.00229) shown in the mosaic plot.

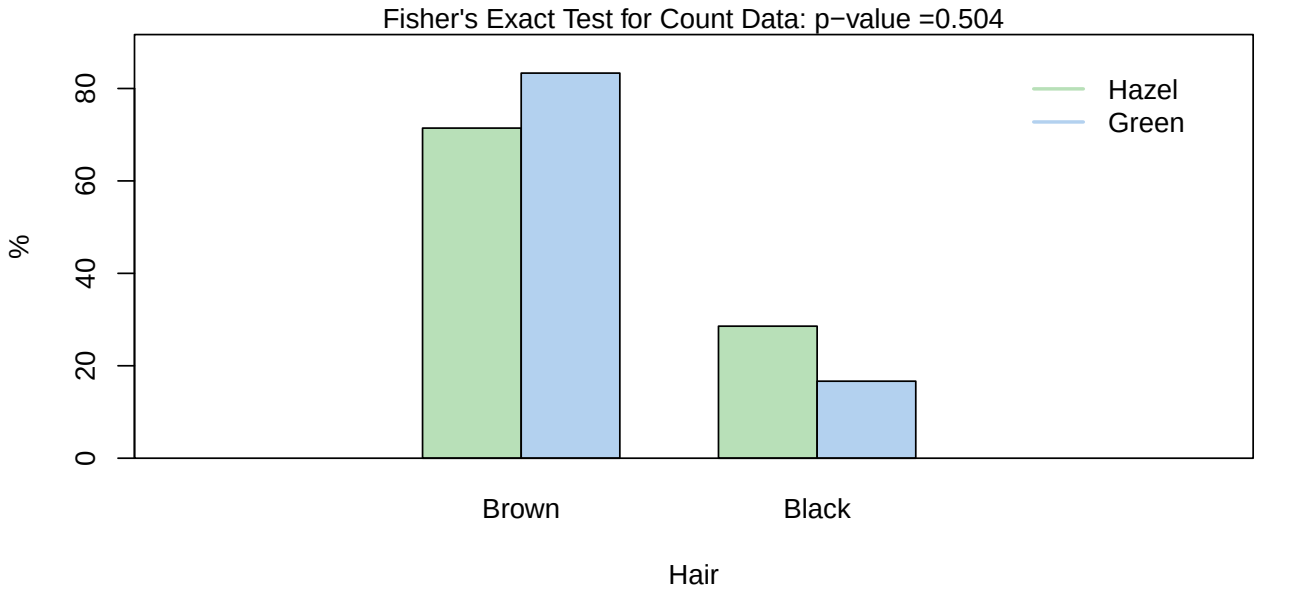
5.7.3 Fisher's exact test (`fisher.test()`)

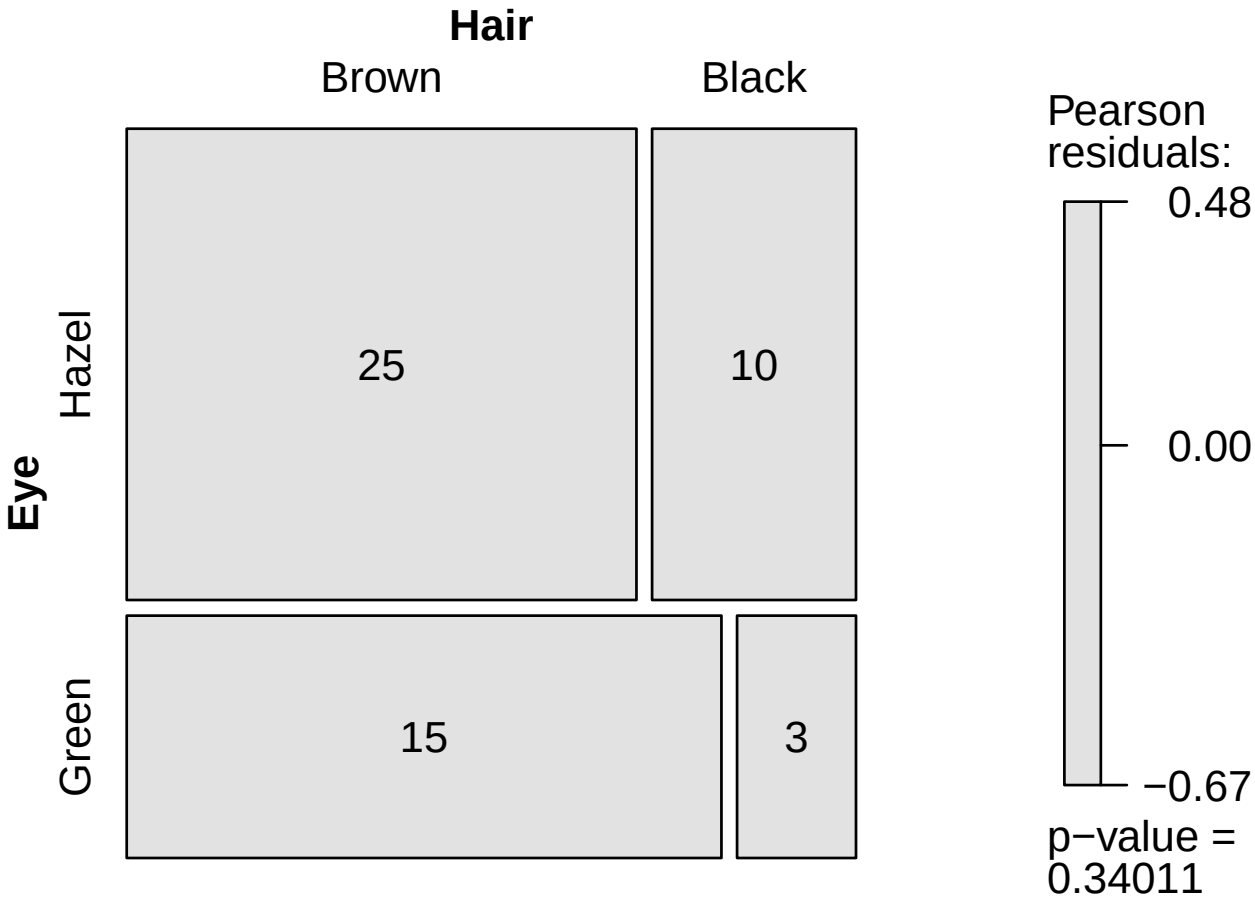
Again, we extract a 2-by-2 contingency table from the full dataset, this time keeping only male participants with black or brown hair and hazel or green eyes.

Pearson's χ^2 test applied to this table would yield an expected frequency less than 5 in one of the four cells (25% of all cells), which violates the requirement that at least 80% of the expected frequencies must be 5 or greater (Cochran 1954).

Therefore, `visstat()` automatically selects Fisher's exact test instead.

```
hair_eye_colour_male <- HairEyeColor[, , 1]
# Slice out a 2 by 2 contingency table
black_brown_hazel_green_male <- hair_eye_colour_male[1:2, 3:4]
# Transform to data frame
black_brown_hazel_green_male <- counts_to_cases(as.data.frame(black_brown_hazel_green_male))
# Fisher test
fisher_stats <- visstat(black_brown_hazel_green_male, "Hair", "Eye")
```





5.8 Saving the graphical output

All generated graphics can be saved in any file format supported by `Cairo()`, including "png", "jpeg", "pdf", "svg", "ps", and "tiff" in the user specified `plotDirectory`. In the following example, we store the graphics in `png` format in the `plotDirectory` `tempdir()`. The file names reflect the statistical test used and the variable names involved.

```
#Graphical output written to plotDirectory: In this example
# a bar chart to visualise the Chi-squared test and mosaic plot showing
#  Pearson's residuals.
#chi_squared_or_fisher_Hair_Eye.png and  mosaic_complete_Hair_Eye.png
visstat(black_brown_hazel_green_male, "Hair", "Eye",
  graphicsoutput = "png", plotDirectory = tempdir())
```

Remove the graphical output from `plotDirectory`:

```
file.remove(file.path(tempdir(), "chi_squared_or_fisher_Hair_Eye.png"))
file.remove(file.path(tempdir(), "mosaic_complete_Hair_Eye.png"))
```

6 Implemented tests

6.1 Numerical response and categorical feature

When the response is numerical and the feature is categorical, test of central tendencies are selected:

`t.test()`, `wilcox.test()`, `aov()`, `oneway.test()`,`kruskal.test()`

6.1.1 Normality assumption check

`shapiro.test()` and `ad.test()`

6.1.2 Homoscedasticity assumption check

`bartlett.test()`

6.1.3 Post-hoc tests

- `TukeyHSD()` (for `aov()`and `oneway.test()`)
- `pairwise.wilcox.test()` (for `kruskal.test()`)

6.2 Numerical response and numerical feature

When both the response and feature are numerical, a simple linear regression model is fitted:

`lm()`

6.3 Categorical response and categorical feature

When both variables are categorical, `visstat()` tests the null hypothesis of independence using one of the following:

- `chisq.test()` (default for larger samples)
- `fisher.test()` (used for small expected cell counts based on Cochran's rule)

Bibliography

Cochran, William G. 1954. "The Combination of Estimates from Different Experiments." *Biometrics* 10 (1): 101. https://doi.org/10.2307/3001666.

Delacre, Marie, Daniël Lakens, and Christophe Leys. 2017. "Why Psychologists Should by Default Use Welch's t-Test Instead of Student's t-Test." *International Review of Social Psychology* 30 (1): 92–101. https://doi.org/10.5334/irsp.82.

Fisher, R. a. 1935. *The Design Of Experiments*.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70. https://www.jstor.org/stable/4615733.

Kruskal, William H., and W. Allen Wallis. 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260): 583–621. https://doi.org/10.2307/2280779.

Lumley, Thomas, Paula Diehr, Scott Emerson, and Lu Chen. 2002. "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health* 23: 151–69. https://doi.org/10.1146/annurev.publhealth.23.100901.140546.

Mann, Henry B., and Donald R. Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *The Annals of Mathematical Statistics* 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

Moser, B K, and G. R. Stevens. 1992. "Homogeneity of Variance in the Two-Sample Means Test." *The American Statistician*, February, 19–21. https://doi.org/10.1080/00031305.1992.10475839.

Pearson, Karl. 1900. "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." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (302): 157–75. https://doi.org/10.1080/14786440009463897.

Rasch, Dieter, Klaus D. Kubinger, and Karl Moder. 2011. "The Two-Sample t Test: Pre-Testing Its Assumptions Does Not Pay Off." *Statistical Papers* 52 (1): 219–31. https://doi.org/10.1007/s00362-009-0224-x.

Satterthwaite, F. E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2 (6): 110–14. https://doi.org/10.2307/3002019.

Šidák, Zbyněk. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62 (318): 626–33. https://doi.org/10.1080/01621459.1967.10482935.

Welch, B. L. 1947. "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved." *Biometrika* 34 (1–2): 28–35. https://doi.org/10.1093/biomet/34.1-2.28.

———. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38 (3/4): 330–36. https://doi.org/10.2307/2332579.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1 (6): 80–83. https://doi.org/10.2307/3001968.

Yates, F. 1934. "Contingency Tables Involving Small Numbers and the $X2$ Test." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1 (2): 217–35. https://doi.org/10.2307/2983604.